Р. КУРАНТ Г. РОББИНС

Что такое математика?

(*Элементарный очерк идей и методов*)

Перевод с английского
под редакцией А. Н. Колмогорова

МЦНМО, 2000

УДК 51(07)
К93
ББК 22.1

What is
Mathematics?

AN ELEMENTARY APPROACH TO
IDEAS AND METHODS
by
RICHARD COURANT
and
HERBERT ROBBINS

Oxford University Press
London – New York – Toronto

Федеральная Программа
Книгоиздания России

Р. Курант, Г. Роббинс

К93 Что такое математика? — 3-е изд., испр. и доп. — М.: МЦНМО, 2001. — 568 с.

ISBN 5–900916–45–6

Книга написана крупным математиком Рихардом Курантом в соавторстве с Гербертом Роббинсом. Она призвана сократить разрыв между математикой, которая преподается в школе, и наиболее живыми и важными для естествознания и техники разделами современной математической науки. Начиная с элементарных понятий, читатель движется к важным областям современной науки. Книга написана доступным языком и является классикой популярного жанра в математике.

Книга предназначена для школьников, студентов, преподавателей, а также для всех интересующихся развитием математики и ее структурой.

ББК 22.1

ISBN 5–900916–45–6

Оглавление

Предисловие к третьему изданию на русском языке

Книга, которую держит в руках читатель,— одно из самых замечательных введений в математику в ряду тех, что обращены к широкой читательской аудитории. Ее замысел выражен в предисловии: «Нет ничего невозможного в том, чтобы, начиная от первооснов и следуя по прямому пути, добраться до таких возвышенных точек, с которых можно ясно обозреть самую сущность и движущие силы современной математики.»

Первый из авторов книги — Рихард Курант (1888–1972) — один из ведущих математиков XX века, ученик Д. Гильберта, иностранный член Академии Наук СССР. Книги Куранта неоднократно издавались на русском языке. На них выросло не одно поколение математиков. Его книги «Уравнения математической физики», «Теория функций», «Уравнения в частных производных», и «Принцип Дирихле» до сих пор остаются основополагающими при изучении математики.

Данную книгу Курант задумал написать в драматический период истории, осенью 1939 г., когда разразилась вторая мировая война. Пятью годами раньше он оказался в Соединенных Штатах Америки, изгнанный фашистами со своей родины — Германии, где он работал в математическом интститутe Гёттингенского университета. Нельзя не отметить огромную заслугу Куранта как организатора в том, что этот институт стал мировым математическим центром. Собственно говоря, Курант, воплотив давнюю мечту Феликса Клейна, основал этот институт. В США Курант создал еще один выдающийся институт (ныне известный как «курантовский институт»), который играл и играет важную роль в развитии прикладной математики во всем мире.

Для осуществления своего замысла — написать книгу, читая которую можно было бы «войти в соприкосновение с самим *содержанием* живой математической науки»,— Курант привлек молодого двадцатичетырехлетнего тополога Герберта Роббинса. Курант, используя свой талант организатора, сумел добыть в те трудные годы немалые материальные средства для издания такого объемного труда. Он долго колебался, выбирая название для своей книги, и окончательно утвердился в нем, лишь поговорив с великим немецким писателем, также лишенным родины, Томасом Манном.

Книга Куранта и Роббинса была переведена на русский язык и подготовлена к печати в 1947 г. Это было очень трудное время для нашей страны. Только что закончилась Великая Отечественная война, потребовавшая немыслимого напряжения. Но, несмотря на это, целесообразность издания труда Куранта и Роббинса была совершенно несомненной для проницательных ученых, думавших о будущем страны.

Однако для того, чтобы книга вышла в свет, потребовалось преодолеть существенные препятствия: у нас началась борьба с космополитизмом, когда русская культура противопоставлялась мировой, а значение последней принижалось. Для выхода книги потребовалось предисловие «От издательства». Оно было вклеено в каждый экземпляр отпечатанного тиража (15 000 экземпляров), между десятой и одиннадцатой страницами, без номеров страниц и без указания о нем в оглавлении.

Требовались особые аргументы для того, чтобы уже напечатанный тираж не был уничтожен. Предисловие было написано Андреем Николаевичем Колмогоровым — одним из величайших математиков уходящего века, хотя и не было подписано им.

Это предисловие — примечательный исторический документ, в котором отражены драматические перипетии того времени. Оно напечатано в добавлении к этому изданию, но мне хочется привести здесь некоторые фрагменты из него о значении книги Куранта и Роббинса. Они актуальны и в наше время, когда живо обсуждаются проблемы математического образования.

Первые три абзаца предисловия обращены к тем основным группам молодежи, для которых, по мнению Колмогорова, книга может быть наиболее полезна. Прежде всего, это школьники, ибо «существует большой разрыв между математикой, которая преподается в средней школе, и наиболее живыми и важными для естествознания и техники разделами современной математической науки». Затем, это студенты инженерных, химических, биологических и сельскохозяйственных вузов, в которых «оставляют совершенно в стороне ряд более общих и новых идей математики... Между тем, эти идеи становятся все более существенными для всей совокупности точных и технических наук». Наконец, это «молодежь, избравшая своей специальностью математику или те разделы естественных наук (механика, астрономия, физика), изучение которых связано с прохождением вполне современного курса математики... [и которая] часто нуждается в том, чтобы еще на стадии перехода из средней школы в высшую в более легкой и наглядной форме познакомиться с различными разделами математики, вплоть до самых важных и современных».

Труд Куранта и Роббинса удовлетворяет потребности этих групп молодежи. Но не только. Эта книга интересна всякому человеку, которому небезразлична судьба научного знания. Вне всякого сомнения, она вхо-

дит в золотой фонд литературы по математике. Книга была переведена на многие языки и сразу же после ее издания стала математическим бестселлером.

* * *

Эта книга была написана шестьдесят лет назад. С тех пор во всем мире и в математической науке произошли весьма значительные изменения. Структура книги Куранта и Роббинса во многом соответствует структуре математики, сложившейся в начале века. Представление об этой структуре дает список основных секций на Втором математическом конгрессе (Париж, 1900 г.): арифметики и алгебры, геометрии, анализа, механики и математической физики. Ныне, в дополнение к этим четырем секциям, на современных конгрессах работают секции математической логики и оснований математики, топологии, алгебраической геометрии, комплексного анализа, теории групп Ли и теории представлений, теории функций и функционального анализа, дифференциальных уравнений с частными производными, обыкновенных дифференциальных уравнений, численных методов, дискретной математики и комбинаторики, теории информации и приложений математики к нефизическим наукам.

Масштаб произошедших изменений не даёт возможности в коротких редакторских примечаниях отразить содержательно достижения в математике за последние две трети века. Поэтому мы ограничились лишь самыми необходимыми комментариями к тексту книги, но при этом значительно пересмотрели и расширили список литературы, включив в него наиболее интересные книги, ориентированные на школьников, вышедшие за последние тридцать лет,

В добавлении помещен также фрагмент книги К. Рид «Курант в Гёттингене и Нью-Йорке», посвященный истории создания книги Куранта и Роббинса.

В. М. Тихомиров

Предисловие ко второму изданию на русском языке

Книга Р. Куранта и Г. Роббинса уже издавалась в СССР в 1947 г. Она пользуется большим успехом у любителей математики самых различных возрастов и уровней подготовки, но давно уже стала библиографической редкостью. В серии «Математическое просвещение» она займет свое почетное место.

Перевод, выполненный для первого издания под редакций покойного проф. В. Л. Гончарова, был выправлен и пополнен по последним английскому (1948) и немецкому (1962) изданиям. Восстановлен также предметный указатель. Список «рекомендованной литературы» следует

оригиналу лишь в части книг, переведенных на русский язык; редакторы русского издания дополнили его рядом книг, имеющихся на русском языке.

Примечания редакторов русского издания немногочисленны (они помечены цифрами, в то время как примечания авторов обозначены звездочками[1].) Редакторы, не желая нарушать цельный и впечатляющий стиль книги, не стремились исправлять и дополнять довольно случайный выбор их указаний на историю вопроса и принадлженость отдельных результатов определенным лицам.

Мы рады поблагодарить проф. Р. Куранта за любезное внимание, оказанное им новому изданию книги на русском языке. В своем коротком обращении к русскому читателю он еще раз подчеркивает руководящую идею своей педагогической деятельности: пропаганду органического единства математики и ее неразрывной связи с естествознанием и техникой. При этом имеется в виду не нравоучения об обязанности математиков быть полезными, а наглядная демонстрация того, что живые источники математического творчества неотделимы от интереса к познанию природы и задачам управления природными явлениями.

В новом издании использованы замечания проф. К. Л. Зигеля и проф. Отто Нейгебауэра, которым мы вместе с авторами выражаем искреннюю признательность.

Москва, *А. Н. Колмогоров*
12 ноября 1966 г.

[1] В настоящем издании не сохранилось. — *Прим. ред. наст. изд.*

К русскому читателю

Выход в свет второго русского издания нашей книги — весьма приятное для меня событие. Я всегда с глубоким восхищением относился к замечательному вкладу в нашу науку, сделанному многими выдающимися математиками Советского Союза. Пожалуй, в большей степени, чем в некоторых странах Запада, русская математическая традиция сохранила идеал единства науки и способствовала упрочению роли математики в научных и технических приложениях. На меня также производит сильнейшее впечатление активное участие, которое принимают крупные математики Советской России в деле подъема математического образования. Я рад, что свое место в русской научно-педагогической литературе по математике заняла и наша книга.

Настоящее издание отличается от предыдущих английских и немецких изданий небольшими исправлениями и уточнениями, которыми мы обязаны, в частности, профессору К. Л. Зигелю, Отто Нейгенбауэру и другим своим коллегам.

Нью-Йоркский университет, *Р. Курант*
9 мая 1966 г.

ПОСВЯЩАЕТСЯ

Эрнсту, Гертруде, Гансу и Леоноре Курант

Предисловие к первому изданию

На протяжении двух с лишним тысячелетий обладание некоторыми, не слишком поверхностными, знаниями в области математики являлось необходимой составной частью интеллектуального багажа каждого образованного человека. В наши дни установленному традицией воспитательному значению математики угрожает серьезная опасность. К сожалению, профессиональные представители математической науки в данном случае не свободны от ответственности. Обучение математике нередко приобретало характер стереотипных упражнений в решении задач шаблонного содержания, что, может быть, и вело к развитию кое-каких формальных навыков, но не призывало к глубокому проникновению в изучаемый предмет и не способствовало развитию подлинной свободы мысли. Научные исследования обнаруживали тенденцию в сторону чрезмерной абстракции и специализации. Приложениям и взаимоотношениям с иными областями не уделялось достаточно внимания. И все же эти малоблагоприятные предпосылки ни в какой мере не могут послужить оправданием для политики сдачи позиций. Напротив, те, кто умеют понимать значение умственной культуры, не могут не выступить — и уже выступают — на ее защиту. Преподаватели, учащиеся — все, хотя бы и не связанные со школой, образованные люди — требуют не идти по линии наименьшего сопротивления, не складывать оружия, а приступить к конструктивной реформе преподавания. Целью является подлинное понимание существа математики как органического целого и как основы научного мышления и действования.

Несколько блестящих книг биографического и исторического содержания и кое-какие публицистические выступления разбудили в широких кругах, казалось бы, безразличных к математике, на самом деле никогда не угасавший к ней интерес. Но знание не может быть достигнуто с помощью одних лишь косвенных средств. Понимание математики не приобретается только безболезненно развлекательными способами — так же как, например, вы не сможете приобрести музыкальной культуры путем чтения журнальных статей (как бы ярко они ни были написаны), если не научитесь *слушать* внимательно и сосредоточенно. Нельзя обойтись без действенного соприкосновения с самим *содержанием* живой

математической науки. С другой стороны, следовало бы избегать всего слишком технического или искусственного, делая изложение математики в одинаковой степени свободным от духа школьной рутины и от мертвящего догматизма, отказывающегося от мотивировок и указания целей, — того самого догматизма, который представляет собой столь неприятное препятствие для честного усилия. Нет ничего невозможного в том, чтобы, начиная от первооснов и следуя по прямому пути, добраться до таких возвышенных точек, с которых можно ясно обозреть самую сущность и движущие силы современной математики.

Настоящая книга делает такую именно попытку. Поскольку она не предполагает иных сведений, кроме тех, которые сообщаются в хорошем школьном курсе, ее можно было бы назвать популярной. Но она — не уступка опасной тенденции устранить всякое напряжение мысли и упражнение. Она предполагает известный уровень умственной зрелости и *готовность усваивать предлагаемое рассуждение.* Книга написана для начинающих и для научных работников, для учащихся и для учителей, для философов и для инженеров; она может быть использована как учебное пособие в учебных заведениях и в библиотеках. Может быть, намерение обратиться к такому широкому кругу читателей является чересчур смелым и самонадеянным. Нужно признать, что под давлением иной работы мы вынуждены были при публикации этой книги искать компромиссы: подготовка велась многие годы, но так и не была по-настоящему закончена. Мы будем рады критике и готовы выслушать пожелания.

Если ответственность за план и философское содержание этой публикации ложится на нижеподписавшегося, то воздаяние ее достоинствам (если таковые имеются) мне подобает разделить с Гербертом Роббинсом. С самого момента присоединения к задуманной работе он отдался ей с увлечением, как своей собственной, и его сотрудничество сыграло решающую роль в окончательном придании книге ее настоящей формы.

Я должен выразить свою глубокую благодарность за помощь многочисленным друзьям. Беседы с Нильсом Бором, Куртом Фридрихом и Отто Нейгебауэром оказали влияние на мои позиции в вопросах философского и исторического характера. Большое количество конструктивных критических замечаний с точки зрения педагога высказала Эдна Крамер. Давид Гильбарг записал лекции, положенные затем в основу книги. Эрнест Курант, Норман Девидс, Чарльз де Прима, Альфред Горн, Герберт Минтцер, Вольфганг Вазов и другие помогли в поистине бесконечной работе по перепечатке рукописи и внесли в нее множество улучшений. Доналд Флендерс внес много ценных предложений и тщательно выверил рукопись к печати. Джон Кнудсен, Герта фон Гумпенберг, Ирвинг Риттер и Отто Нейгебауэр изготовили чертежи. Часть упражнений для приложения в конце книги исходит от Г. Уитни. Кур-

сы лекций и статьи, положенные в основу книги, были осуществлены благодаря щедрой поддержке Отдела народного образования Рокфеллеровского фонда. Я должен также поблагодарить издательство Waverly Press, особенно г-на Гровера К. Орта, за чрезвычайно квалифицированную работу и издательство Oxford University Press, особенно г-на Филипа Водрена и г-на У. Омана, за инициативу и поддержку.

Нью-Рошель, Нью-Йорк, *Р. Курант*
22 августа 1941 г.

Предисловие ко второму, третьему и четвертому изданиям

В последний год, под воздействием совершающихся событий, возник усиленный спрос на математическую информацию и соответствующий инструктивный материал. Сейчас больше чем когда-либо существует опасность выхолащивания и разочарований, если только учащиеся (и учителя) не сумеют увидеть и схватить то, что лежит за формулами и преобразованиями, — истинное существо и содержание математики. Именно для тех, кто видит глубже, была написана эта книга, и отклики на первое издание поддерживают в авторах надежду, что она принесет пользу.

Благодарим читателей, чей критические замечания позволили внести в новые издания многочисленные поправки и улучшения. За большую помощь в подготовке четвертого издания сердечно благодарим г-жу Наташу Артин.

Нью-Рошель, Нью-Йорк, *Р. Курант*
18 марта 1943 г.
10 октября 1945 г.
28 октября 1947 г.

Как пользоваться книгой

Порядок изложения в книге — систематический, но это не значит никоим образом, что читатель обязан читать ее подряд — страницу за страницей, главу за главой. Главы в значительной степени независимы одна от другой. Часто начало раздела покажется легкодоступным, но потом дорога постепенно пойдет вверх, становясь круче в конце главы и в дополнениях к ней. Поэтому читатель, нуждающийся скорее в общей информации, чем в приобретении специальных знаний, поступит правильно, если удовлетворится таким отбором материала, который может быть осуществлен по принципу избегания более детализированных рассмотрений.

Учащийся с ограниченной математической подготовкой пусть выбирает по своему вкусу. Звездочками и мелким шрифтом отмечено то, что может быть опущено при первом чтении без серьезного ущерба для понимания последующего. Больше того, беды не будет, если при изучении книги читатель ограничится теми разделами или главами, которые представляют для него наибольший интерес. Большинство упражнений не носит чисто формального характера; более трудные отмечены звездочкой. Не надо слишком огорчаться, если вы не сумеете выполнить некоторые из них.

Преподаватели школ найдут в главах, посвященных геометрическим построениям и максимумам и минимумам, материал, подходящий для кружковых занятий или для отдельных групп учащихся.

Мы надеемся, что книга сможет послужить и учащимся разных классов колледжей и лицам тех или иных профессий, действительно интересующимся проблемами точного знания. Она может быть положена в основу «свободных» курсов в колледжах по основным понятиям математики. Главы III, IV и V подходят для курса геометрии, тогда как главы VI и VIII, вместе взятые, образуют законченное изложение основ анализа с опорой скорее на понимание, чем на достижение технического совершенства. Они могут быть использованы в качестве вводного текста преподавателем, который пожелал бы дополнить учебный курс в соответствии с теми или иными специфическими потребностями, и в особенности — обогатить его более разнообразными примерами. Многочисленные упражнения разбросаны по всей книге; дополнительное собрание упражнений в конце могло бы, как мы полагаем, облегчить ее использование в школьной обстановке.

Мы надеемся, что и специалист обнаружит кое-что интересное в деталях и в иных элементарных рассуждениях, содержащих в себе зерно более широких идей.

Что такое математика?

Математика содержит в себе черты волевой деятельности, умозрительного рассуждения и стремления к эстетическому совершенству. Ее основные и взаимно противоположные элементы — логика и интуиция, анализ и конструкция, общность и конкретность. Как бы ни были различны точки зрения, питаемые теми или иными традициями, только совместное действие этих полярных начал и борьба за их синтез обеспечивают жизненность, полезность и высокую ценность математической науки.

Без сомнения, движение вперед в области математики обусловлено возникновением потребностей, в большей или меньшей мере носящих практический характер. Но раз возникшее, оно неизбежно приобретает внутренний размах и выходит за границы непосредственной полезности. Совершающееся таким образом превращение прикладной науки в теоретическую наблюдается в истории древности, но не в меньшей степени также и в наши дни: достаточно принять во внимание тот вклад, который сделан в современную математику инженерами и физиками.

Самые ранние из дошедших до нас образцов математической мысли появились на Востоке: около двух тысячелетий до нашей эры вавилоняне собрали обширный материал, который мы склонны были бы в настоящее время отнести к элементарной алгебре. Но как наука в современном смысле слова математика возникает позднее на греческой почве, в пятом и четвертом столетиях до нашей эры. Все усиливающееся соприкосновение между Востоком и Грецией, начавшееся во времена Персидской империи и достигшее апогея в период, непосредственно следующий за экспедициями Александра Македонского, обеспечило грекам возможность перенять достижения вавилонян в области математики и астрономии. Математика не замедлила стать объектом философских дискуссий, обычных в греческих городах-государствах. Таким образом, греческие мыслители осознали значительные трудности, связанные с основными математическими концепциями — непрерывностью, движением, бесконечностью — и с проблемой измерения произвольных величин данными заранее единицами. Но обнаружилась и решимость преодолеть препятствия: возникшая в результате великолепного усилия мысли евдоксова теория геометрического континуума представляет собой такое достижение, которое можно поставить в один ряд только с современной теорией иррациональных чисел. От Евдокса идет аксиоматико-дедуктивное напрвление в математике, проявившееся вполне отчетливо в «Началах» Евклида.

Хотя теоретико-постулативная тенденция незыблемо остается одной из самых ярких особенностей греческой математики и, как таковая, оказала беспримерное влияние на дальнейшее развитие науки, тем не

менее необходимо со всей энергией указать, что практические потребности и связь с физической реальностью участвовали никак не в меньшей мере в создании античной математики и что изложению, свободному от евклидовой строгости, очень часто отдавалось предпочтение.

Не исключено, что именно слишком раннее открытие трудностей, связанных с «несоизмеримыми» величинами, помешало грекам развить искусство численных операций, сделавшее в предшествовавшие эпохи значительные успехи на Востоке. Вместо этого они стали искать пути в дебрях чистой аксиоматической геометрии. Так началось одно из странных блужданий в истории науки, и, может быть, были при этом упущены блестящие возможности. Почти на два тысячелетия авторитет греческой геометрической традиции задержал неизбежную эволюцию идеи числа и буквенного исчисления, положенных впоследствии в основу точных наук.

После периода медленного накопления сил — с возникновением в XVII столетии аналитической геометрии и дифференциального и интегрального исчислений — открылась бурная революционная фаза в развитии математики и физики. В XVII и XVIII вв. греческий идеал аксиоматической кристаллизации и систематической дедукции потускнел и утерял свое влияние, хотя античная геометрия продолжала высоко расцениваться. Логически безупречное мышление, отправляющееся от отчетливых определений и «очевидных», взаимно не противоречащих аксиом, перестало импонировать новым пионерам математического знания. Предавшись подлинной оргии интуитивных догадок, перемешивая неоспоримые заключения с бессмысленными полумистическими утверждениями, слепо доверяясь сверхчеловеческой силе формальных процедур, они открыли новый математический мир, полный несметных богатств. Но мало-помалу экстатическое состояние мысли, упоенной головокружительными успехами, уступило место духу сдержанности и критицизма. В XIX столетии осознание необходимости консолидировать науку, особенно в связи с нуждами высшего образования, после Французской революции получившего широкое распространение, повело к ревизии основ новой математики; в частности, внимание было направлено к дифференциальному и интегральному исчислениям и к уяснению подразумеваемого анализом понятия предела. Таким образом, XIX век не только стал эпохой новых успехов, но и был ознаменован плодотворным возвратом к классическому идеалу точности и строгости доказательств. В этом отношении греческий образец был даже превзойден. Еще один раз маятник качнулся в сторону логической безупречности и отвлеченности. В настоящее время мы еще, по-видимому, не вышли из этого периода, хотя позволительно надеяться, что установившийся прискорбный разрыв между чистой математикой и ее жизненными приложениями, неизбежный, по-видимому, во времена критических ревизий, сменится эрой более тесного единения. Приобре-

тенный запас внутренних сил и, помимо всего прочего, чрезвычайное упрощение, достигаемое на основе ясного понимания, позволяют сегодня манипулировать математической теорией таким образом, чтобы приложения не упускались из виду. Установить еще раз органическую связь между чистым и прикладным знанием, здоровое равновесие между абстрактной общностью и полнокровной конкретностью — вот как нам представляется задача математики в непосредственно обозримом будущем.

Здесь не место входить в подробный философский или психологический анализ математики. Хочется отметить все же некоторые моменты. Чрезмерное подчеркивание аксиоматико-дедуктивного характера математики представляется мне весьма опасным. Конечно, начало конструктивного творчества, интуитивное начало, являющееся источником наших идей и доводов в их пользу, с трудом укладываются в простые философские формулировки; и тем не менее именно это начало есть подлинная суть любого математического открытия, даже если оно относится к самым абстрактным областям. Если целью и является четкая дедуктивная форма, то движущая сила математики — это интуиция и конструкции. В допущении, что математика есть не более чем система следствий, извлекаемых из определений и постулатов, которые должны быть только совместимы между собой, а в остальном являются продуктом свободной фантазии математиков, таится серьезная угроза для самого существования науки. Если бы это было действительно так, математика была бы занятием, недостойным мыслящего человека. Она была бы просто игрой с определениями, правилами и силлогизмами, не имеющей ни причины, ни цели. Представление, согласно которому человеческий интеллект может творить лишенные какого бы то ни было смысла системы постулатов, есть обман, точнее, полуправда.

Получать результаты, имеющие научную ценность, свободный разум может, только подчиняясь суровой ответственности перед природой, только следуя некоей внутренней необходимости.

Хотя созерцательное направление логического анализа и не представляет всей математики, оно способствовало более глубокому пониманию математических фактов и их взаимозависимости и более ясному овладению существом математических понятий. Именно из этого направления выросла современная точка зрения на математику как на образец универсально приложимого научного метода.

Каких бы философских позиций мы ни придерживались, все задачи научного исследования сводятся к нашему отношению к воспринимаемым объектам и инструментам исследования. Конечно, восприятие само по себе еще не есть ни знание, ни понимание; нужно еще согласовать их между собой и истолковать в терминах некоторых лежащих за ними сущностей, «вещей в себе», не являющихся предметами непосредственно

физического изучения, а принадлежащими к метафизической сфере. Но для научного метода существенным является отказ от метафизических умозрений и, в конечном счете, представление всех наблюдаемых фактов в форме понятий и конструкций. Отказ от претензии понимания природы «вещей в себе», от постижения «окончательной истины», от разгадки внутренней сущности мира, быть может, будет психологически тягостен для наивных энтузиастов, но на самом-то деле этот отказ оказался в высшей степени плодотворным для развития современной научной мысли.

Некоторым из величайших открытий физики мы обязаны смелому следованию принципу устранения метафизики. Когда Эйнштейн попытался свести понятие «одновременных событий, происходящих в разных местах» к наблюдаемым явлениям, когда он понял, что вера в то, что это понятие само по себе непременно должно иметь какой-то точный смысл, есть попросту метафизический предрассудок, в этом открытии уже было заключено ядро его теории относительности. Когда Нильс Бор и его ученики вдумались в тот факт, что любое физическое наблюдение связано с взаимодействием между прибором и наблюдаемым объектом, то им стало ясно, что точное одновременное определение положения и скорости частицы в том смысле, в каком это понимается в физике, невозможно. Далеко идущие следствия этого открытия, составившие современную систему квантовой механики, хорошо известны ныне каждому физику. В XIX столетии господствовала идея, согласно которой механические силы и передвижения частиц в пространстве суть вещи в себе, а электричество, свет и магнетизм можно свести к механическим явлениям (или «объяснить» в механических терминах), подобно тому как это было сделано с теорией теплоты. Была выдвинута концепция гипотетической среды — так называемого «эфира»,— способной к не вполне понятным механическим передвижениям, представляющимся нам в качестве света или электричества. Постепенно выяснилось, что этот эфир принципиально ненаблюдаем, т. е. что это понятие принадлежит скорее метафизике, нежели физике. Вначале с сожалением, а затем с облегчением идея механического объяснения световых и электрических явлений — а вместе с ней и понятие эфира — была окончательно отброшена.

Подобная же ситуация, и даже еще более отчетливая, создалась и в математике. В течение столетий математики рассматривали интересующие их объекты — числа, прямые и т. д. — как некие субстанции, вещи в себе. Поскольку, однако, эти «сущности» упорно не поддавались попыткам точного описания их природы, математики девятнадцатого столетия стали понемногу укрепляться в мысли, что вопрос о значении этих понятий как субстанциальных объектов в рамках математики (да и где бы то ни было) просто не имеет смысла. Математические утверждения, в которые входят эти термины, относятся вовсе не к физической ре-

альности; они лишь устанавливают взаимосвязи между математически «неопределимыми объектами» и правила оперирования с ними. Вопрос о том, чем «на самом деле» *являются* точки, прямые и числа, не может и не должна обсуждать математическая наука. Действительно существенными и имеющими непосредственное касательство к «проверяемым» фактам являются структура и взаимосвязи между этими объектами: что две точки определяют прямую, что из чисел по определенным правилам получаются другие числа, и т. п. Ясное осознание необходимости отказа от представления об основных математических понятиях как о реально существующих предметах явилось одним из самых важных и плодотворных завоеваний современного аксиоматического развития математики.

К счастью, творческая мысль забывает о догматических философских верованиях, как только привязанность к ним становится на пути конструктивных открытий. И для специалистов, и для любителей не философия, а именно активные занятия самой математикой смогут дать ответ на вопрос: Что такое математика?

ГЛАВА I

Натуральные числа

Введение

Число — это основное понятие современной математики. Но что такое число? Если мы говорим, что $\frac{1}{2}+\frac{1}{2}=1$, $\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$ или что $(-1)\cdot(-1)=1$, то какой смысл вкладывается в эти утверждения? В школе мы изучаем технику действий с дробями и с отрицательными числами, но, чтобы приобрести подлинное понимание того, как устроена система чисел, недостаточно ограничиваться элементарными сведениями и нужно пойти несколько дальше. Греки в древнее время в основу созданной ими математики положили геометрические концепции точки и прямой; руководящим принципом современной математики стало сведение в конечном счете всех утверждений к утверждениям, касающимся *натуральных чисел* 1, 2, 3, ... «Бог создал натуральные числа, все прочее — дело рук человека». Этими словами Леопольд Кронекер (1823–1891) определил тот прочный фундамент, на котором может быть построено здание математики.

Числа служат для того, чтобы считать объекты, входящие в состав тех или иных объединений или собраний. Числа решительно никак не связаны с индивидуальной характеристикой считаемых объектов. Так, число «шесть» есть результат абстрагирования, производимого при рассмотрении всевозможных совокупностей, состоящих из шести предметов: оно нисколько не зависит ни от специфических свойств этих объектов, ни от употребляемых символов (обозначений). Но абстрактный характер идеи числа становится ясным только на очень высокой ступени интеллектуального развития. В глазах детей числа всегда остаются соединенными с самими осязаемыми объектами — допустим, пальцами или камешками; в языках народов числа также трактуются конкретно: для обозначения предметов различных типов употребляются различные сочетания числительных.

Мы воспользуемся тем, что математик (как таковой) не обязан заниматься философской проблемой перехода от совокупностей конкретных предметов к абстрактному понятию числа. Мы примем поэтому натуральные числа как данные вместе с двумя основными операциями, над

ними совершаемыми: сложением и умножением.

§ 1. Операции над целыми числами

1. Законы арифметики. Математическую теорию *натуральных* (иначе, *целых положительных*) *чисел* называют *арифметикой.* Эта теория основана на том факте, что сложение и умножение целых чисел подчинены некоторым законам. Чтобы сформулировать эти законы во всей их общности, нельзя воспользоваться символами вроде 1, 2, 3, относящимися к определенным, конкретным числам. Утверждение

$$1 + 2 = 2 + 1$$

есть только частный случай общего закона, содержание которого заключается в том, что сумма двух чисел не зависит от порядка, в котором мы рассматриваем эти числа. Если мы хотим выразить ту мысль, что некоторое соотношение между целыми числами имеет место (оправдывается, осуществляется), каковы бы ни были рассматриваемые числа, то будем обозначать их символически, т. е. условно, буквами a, b, c, ... Раз такого рода соглашение принято, сформулировать пять основных законов арифметики — очевидно, близко знакомых читателю — не представит труда:

$$1)\ a + b = b + a, \qquad 2)\ ab = ba, \qquad 3)\ a + (b + c) = (a + b) + c,$$
$$4)\ a(bc) = (ab)c, \qquad 5)\ a(b + c) = ab + bc.$$

Два первых закона — *коммутативный* (переместительный) закон сложения и коммутативный закон умножения — говорят, что при сложении и при умножении можно менять порядок чисел, над которыми совершается действие. Третий — *ассоциативный* (сочетательный) закон сложения — гласит, что при сложении трех чисел получается один и тот же результат независимо от того, прибавим ли мы к первому числу сумму второго и третьего, или прибавим третье к сумме первого и второго. Четвертый закон есть ассоциативный закон умножения. Последний — *дистрибутивный* (распределительный) закон — устанавливает то обстоятельство, что при умножении суммы на некоторое целое число можно умножить на это число каждое слагаемое и полученные произведения сложить.

Эти арифметические законы совсем просты и, пожалуй, могут показаться очевидными. Но следует все же заметить, что к иного рода объектам — не к целым числам — они могут оказаться и неприменимыми. Например, если a и b обозначают не числа, а химические вещества и если «сложение» понимается в смысле обычной речи, то легко понять, что коммутативный закон сложения не всегда оправдывается. В самом деле, если, скажем, к воде будем прибавлять серную кислоту, то получится разбавленный раствор, тогда как прибавление воды

к чистой серной кислоте может закончиться неблагополучно для экспериментатора. С помощью таких же иллюстраций можно показать, что в химической «арифметике» иногда нарушаются и ассоциативный, и дистрибутивный законы сложения. Итак, можно вообразить и такие типы арифметических систем, в которых один или несколько законов 1)–5) теряют силу. Такие системы действительно изучались современной математикой. Основа, на которой покоятся законы 1)–5), дается конкретной моделью для абстрактного понятия целого числа. Вместо того чтобы пользоваться обыкновенными знаками 1, 2, 3 и т. д., станем обозначать число предметов в данной совокупности (например, яблок на данном дереве) системой точек в четырехугольном «ящичке» — таким образом, чтобы каждому предмету соответствовало по одной точке. Оперируя этими ящичками, мы сможем исследовать законы арифметики целых чисел. Чтобы сложить два целых числа a и b, мы сдвигаем вместе соответствующие ящички и затем уничтожаем перегородку. Чтобы

[• • • • •]

Рис. 1. Сложение

умножить a на b, мы выстроим точки в двух ящичках в ряд и затем устроим новый ящичек, в котором точки будут расположены так, что образуют a горизонтальных и b вертикальных рядов. И тогда ясно видно, что правила 1)–5) выражают интуитивно очевидные свойства введенных операций с ящичками.

[• • • • •] × [• • • •]

Рис. 2. Умножение

[• • •] × ([• •]

Рис. 3. Дистрибутивный закон

На основе определения сложения двух целых чисел можно теперь дать определение *неравенства*. Каждое из двух эквивалентных утверждений, именно $a < b$ («a меньше, чем b») и $b > a$ («b больше, чем a»), обозначает, что ящичек b может быть получен из ящичка a посредством прибавления надлежащим образом выбранного третьего ящичка c таким



Рис. 4. Вычитание

образом, что $b = a + c$. Если это так, то мы напишем

$$c = b - a,$$

чем и определяется операция *вычитания.*

Сложение и вычитание называют *обратными* операциями, так как если, например, к числу a прибавить число d, а затем из того, что получится, отнять d, то получится снова исходное число a:

$$(a + d) - d = a.$$

Нужно заметить, что число $b - a$ было определено только при условии $b > a$. Значение символа $b - a$ как *отрицательного* целого числа при условии $b < a$ будет рассмотрено далее (стр. 73 и след.). Часто бывает удобно пользоваться обозначением $b \geqslant a$ («b больше или равно a») или $a \leqslant b$ («a меньше или равно b», «a не превосходит b»), понимая под этим не что иное, как отрицание того, что $a > b$. Таким образом, можно написать $2 \geqslant 2$, и можно также написать $3 \geqslant 2$.

Мы можем еще несколько расширить область положительных целых чисел, которые мы изображаем ящичками с точками. Введем целое число *нуль*, изображаемое совершенно пустым ящичком; условимся обозначать такой пустой ящичек обычным символом 0. Тогда, согласно нашему определению сложения и умножения, каково бы ни было целое число a, получаются соотношения

$$a + 0 = a, \qquad a \cdot 0 = 0.$$

Действительно, $a + 0$ обозначает прибавление пустого ящичка к ящичку a, а $a \cdot 0$ обозначает ящичек, в котором вовсе нет вертикальных рядов, т. е. пустой ящичек. Тогда уже вполне естественно расширить определение вычитания, полагая

$$a - a = 0$$

при любом a. Таковы характерные арифметические свойства нуля.

Геометрические модели вроде ящичков с точками (сюда относится древний абак) широко применялись при арифметических вычислениях вплоть до конца средневековья и только мало-помалу уступили место гораздо более совершенным символическим методам, основанным на десятичной системе.

2. Представление целых чисел с помощью письменных знаков (нумерация). Необходимо очень тщательно делать различие между целым числом и тем символом (например, 5, V и т. п.), которым

пользуются для его письменного воспроизведения. В нашей десятичной системе нуль и девять целых натуральных чисел обозначаются цифрами 0, 1, 2, 3, ..., 9. Числа большей величины, как, скажем, «триста семьдесят два», представляются в виде

$$300 + 70 + 2 = 3 \cdot 10^2 + 7 \cdot 10 + 2$$

и в десятичной системе записываются символом 372. Существенно в данном случае то, что смысл каждой из цифр 3, 7, 2 зависит от ее положения — от того, стоит ли она на месте единиц, десятков или сотен. Используя «поместное значение» цифр (позиционный принцип), мы имеем возможность изобразить любое натуральное число с помощью всего лишь десяти цифр в их различных комбинациях. Общее правило такого изображения дается схемой, которая иллюстрируется примером

$$z = a \cdot 10^3 + b \cdot 10^2 + c \cdot 10 + d,$$

где a, b, c, d представляют собой целые числа в пределах от нуля до девяти. Число z в этом случае сокращенно обозначается символом

$$abcd.$$

Заметим, между прочим, что коэффициенты d, c, b, a являются не чем иным, как остатками при последовательном делении числа z на 10. Так, например,

$$\begin{array}{r|l}
372 & 10 \\ \hline
2 & 37 \;|\; 10 \\
 & 7 \quad\; 3 \;|\; 10 \\
 & \qquad 3 \quad\; 0
\end{array}$$

С помощью написанного выше выражения для числа z можно изображать только те числа, которые меньше десяти тысяч, так как числа большие, чем десять тысяч, требуют пяти или большего числа цифр. Если z есть число, заключенное между десятью тысячами и ста тысячами, то его можно представить в виде

$$z = a \cdot 10^4 + b \cdot 10^3 + c \cdot 10^2 + d \cdot 10 + e$$

и записать символически как $abcde$. Подобное же утверждение справедливо относительно чисел, заключенных между ста тысячами и одним миллионом, и т. д. Чрезвычайно важно располагать способом, позволяющим высказать результат, к которому мы приходим, во всей его общности посредством одной-единственной формулы. Мы можем добиться этой цели, если обозначим различные коэффициенты e, d, c, ... одной и той же буквой a с различными значками (индексами) a_0, a_1, a_2, ..., а то обстоятельство, что степени числа 10 могут быть сколь угодно большими, выразим тем, что высшую степень числа 10 обозначим не 10^3 или 10^4, как в предыдущих примерах, а станем писать 10^n, понимая

под n совершенно произвольное натуральное число. В таком случае любое целое число z в десятичной системе может быть представлено в виде

$$z = a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \ldots + a_1 \cdot 10 + a_0 \tag{1}$$

и записано посредством символа

$$a_n a_{n-1} a_{n-2} \ldots a_1 a_0.$$

Как и в рассмотренном выше частном примере, мы обнаруживаем, что a_0, a_1, a_2, …, a_n являются остатками при последовательном делении z на 10.

В десятичной системе число «десять» играет особую роль как «основание» системы. Тот, кому приходится встречаться лишь с практическими вычислениями, может не отдавать себе отчета в том, что такое выделение числа десять не является существенным и что роль основания способно было бы играть любое целое число, большее единицы. Например, была бы вполне возможна семеричная система с основанием семь. В такой системе целое число представлялось бы в виде

$$b_n \cdot 7^n + b_{n-1} \cdot 7^{n-1} + \ldots + b_1 \cdot 7 + b_0, \tag{2}$$

где коэффициенты b обозначают числа в пределах от нуля до шести, и оно записывалось бы посредством символа

$$b_n b_{n-1} \ldots b_1 b_0.$$

Так, число «сто девять» в семеричной системе обозначалось бы символом 214, потому что

$$109 = 2 \cdot 7^2 + 1 \cdot 7 + 4.$$

В качестве упражнения читатель может вывести общее правило для перехода от основания 10 к любому основанию B: нужно выполнять последовательные деления на B, начиная с данного числа z; остатки и будут «цифрами» при записи числа в системе с основанием B. Например,

$$\begin{array}{rrrr}
109 & \lfloor 7 & & \\
4 & 15 & \lfloor 7 & \\
 & 1 & 2 & \lfloor 7 \\
 & & 2 & 0
\end{array}$$

109 (в десятичной системе) = 214 (в семеричной системе).

Естественно, возникает вопрос: не был ли бы особенно желательным выбор какого-либо специального числа в качестве основания системы счисления? Мы увидим дальше, что слишком маленькое основание должно было бы вызвать кое-какие неудобства; с другой стороны, слишком большое основание потребовало бы заучивания многих цифр и знания расширенной таблицы умножения. Высказывались соображения в пользу системы с основанием 12 («двенадцатиричной»): указывалось,

что 12 делится без остатка на два, на три, на четыре и на шесть, и потому вычисления, связанные с делениями и дробями, при основании 12 были бы несколько проще. Чтобы написать произвольное число в двенадцатиричной системе, понадобились бы две лишние цифры — для обозначения чисел «десять» и «одиннадцать». Пусть α обозначало бы десять, а β — одиннадцать. Тогда в двенадцатиричной системе «двенадцать» пришлось бы написать в виде 10, «двадцать два» — в виде 1α, «двадцать три» — в виде 1β, а «сто тридцать один» — в виде $\alpha\beta$.

Изобретение позиционной нумерации, основанной на поместном значении цифр, приписывается шумерийцам и вавилонянам; развита была такая нумерация индусами и имела неоценимые последствия в истории человеческой цивилизации. Более древние системы нумерации были построены исключительно на аддитивном принципе.[1] Так, в римской нумерации CXVII обозначает «сто + десять + пять + один + один + один». Египетская, еврейская и греческая системы были на том же уровне. Неудобством чисто аддитивной системы является то обстоятельство, что с увеличением изображаемых чисел требуется неограниченное число новых символов. Но главнейшим недостатком древних систем (вроде римской) было то, что сама процедура счета была очень трудна: даже самые простые задачи могли решать только специалисты-профессионалы. Совсем иначе обстоит дело с распространенной в наше время индусской «позиционной» системой. В средневековой Европе она появилась через итальянских купцов, в свою очередь заимствовавших ее у мусульман. Позиционная система обладает тем чрезвычайно выгодным свойством, что все числа, и малые и большие, могут быть записаны с помощью небольшого числа различных символов; в десятичной системе таковыми являются «арабские цифры» $0, 1, 2, \ldots, 9$. Не меньшее значение имеет и легкость счета в этой системе. Правила действий с числами, записываемыми по позиционному принципу, могут быть резюмированы в виде таблиц сложения и умножения и могут быть раз навсегда выучены на память. Старинному методу счета, которым раньше владели лишь немногие избранные, теперь обучают разве лишь в начальных школах. В истории культуры найдется немного примеров того, чтобы научный прогресс оказал на практическую жизнь столь глубокое, столь облегчающее влияние.

3. Арифметические действия в недесятичных системах счисления. Исключительная роль десятка восходит к истокам цивилизации и без всякого сомнения связана со счетом по пальцам на двух руках. Но наименования числительных в разных языках указывают и на

[1] На самом деле элементы «позиционности» есть и в римской нумерации, во всяком случае, порядок расположения «разрядов» играет роль; так, $\mathrm{VI} = \mathrm{V} + \mathrm{I}$, но $\mathrm{IV} = \mathrm{V} - \mathrm{I}$, $\mathrm{LX} = \mathrm{L} + \mathrm{X}$, но $\mathrm{XL} = \mathrm{L} - \mathrm{X}$ и т. п. — *Прим. ред.*

наличие (в былые времена) иных систем счисления, именно с основаниями двадцать и двенадцать. В английском и немецком языках слова, обозначающие 11 и 12, построены не по десятичному принципу, сочетающему десятки с единицами: они лингвистически независимы от слов, обозначающих число 10. Во французском языке слова, обозначающие 20 и 80, позволяют предполагать первоначальное существование системы с основанием 20, используемой для тех или иных надобностей. В датском языке слово halvfirsinds-tyve, обозначающее 70, буквально переводится «полпути от трижды двадцать до четырежды двадцать». Вавилонские астрономы пользовались системой, являвшейся отчасти шестидесятеричной (с основанием 60), и именно в этом обстоятельстве следует искать объяснение того факта, что час и угловой градус подразделены на 60 минут.

В недесятичных системах счисления правила арифметики, конечно, те же самые, но таблицы сложения и умножения однозначных чисел отличны от наших десятичных. Будучи приучены к десятичной системе и связаны наименованиями числительных в нашем языке, мы, если попытаемся считать по иным системам, сначала испытаем известное неудобство. Попробуем поупражняться в умножении по семеричной системе. Прежде чем приступить к этому, рекомендуется выписать две таблички, которыми придется пользоваться.

Сложение

	1	2	3	4	5	6
1	2	3	4	5	6	10
2	3	4	5	6	10	11
3	4	5	6	10	11	12
4	5	6	10	11	12	13
5	6	10	11	12	13	14
6	10	11	12	13	14	15

Умножение

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	11	13	15
3	3	6	12	15	21	24
4	4	11	15	22	26	33
5	5	13	21	26	34	42
6	6	15	24	33	42	51

Станем теперь умножать 265 на 24, причем эти числа предполагаются написанными в семеричной системе. (Если написать числа по десятичной системе, то речь идет об умножении 145 на 18.) Начнем с умножения 5 на 4, что, как показывает таблица умножения, дает 26.

$$\begin{array}{r} \times\begin{array}{r}265\\24\end{array} \\ \hline +\begin{array}{r}1456\\563\end{array} \\ \hline 10416 \end{array}$$

Мы пишем 6 на месте единицы, затем переносим двойку в следующий разряд. Далее, находим, что $4 \cdot 6 = 33$ и что $33 + 2 = 35$. Пишем в

произведении 5 и продолжаем таким же образом, пока умножение не закончится. При сложении чисел 1456 и 5630 на месте единиц получаем $6 + 0 = 6$, затем на месте семерок $5 + 3 = 11$. Пишем 1 и 1 переносим на место «сорокадевяток», где получается $1 + 6 + 4 = 14$. Окончательный результат: $265 \cdot 24 = 10416$.

Для проверки проделаем то же действие в десятичной системе. Чтобы переписать число 10 416 по десятичной системе, придется найти степени 7 вплоть до четвертой: $7^2 = 49$, $7^3 = 343$, $7^4 = 2401$. Отсюда следует, что $10416 = 2401 + 4 \cdot 49 + 7 + 6$, причем правая часть равенства записана уже по десятичной системе. Складывая числа, мы находим, что число 10 416, записанное по семеричной системе, равно числу 2610, записанному по десятичной. Умножим теперь 145 на 18 в десятичной системе: получается как раз 2610.

Упражнения. 1) Составьте таблицы сложения и умножения в двенадцатеричной системе и проделайте несколько примеров вроде приведенного выше.

2) Напишите «тридцать» и «сто тридцать три» в системах с основаниями 7, 11, 12.

3) Что обозначают символы 11111 и 21212 в этих системах?

4) Составьте таблицы сложения и умножения для систем с основаниями 5, 11, 13.

С теоретической точки зрения система, построенная по позиционному принципу с основанием 2, выделяется в том смысле, что это основание — наименьшее возможное. В этой *двоичной* (диадической, бинарной) системе имеются лишь две цифры: 0 и 1; всякое иное число записывается как комбинация этих символов. Таблицы сложения и умножения сводятся к двум правилам: $1 + 1 = 10$ и $1 \cdot 1 = 1$. Но непрактичность такой системы достаточно очевидна[1]: чтобы изобразить уже небольшие числа, нужны длинные выражения. Так, число «семьдесят девять», которое представляется в виде $1 \cdot 2^6 + 0 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2 + 1$, записывается в двоичной системе как 1 001 111.

Чтобы проиллюстрировать, насколько просто производится умножение в двоичной системе, перемножим числа семь и пять, которые записываются соответственно в виде 111 и 101. Принимая во внимание,

[1] Со времени написания книги (1941 г.; последнее английское издание, которым мы располагали, вышло в 1948 г.) столь «непрактичная» для обычного счета двоичная система получила широкие и общеизвестные применения в машинной математике (идея которых — «кодирование» любого текста с помощью алфавита из двух знаков — предугадывается в приводимой ниже фразе Лапласа о Лейбнице). — *Прим. ред.*

что в этой системе $1+1=10$, мы пишем:

$$\begin{array}{r} \times\ 111 \\ 101 \\ \hline +\ \ 111 \\ 111 \\ \hline 100011 \end{array} = 2^5 + 2 + 1,$$

и в итоге, как и следовало ожидать, получается тридцать пять.

Готфрид Вильгельм Лейбниц (1646–1716), один из величайших умов своего времени, расценивал двоичную систему чрезвычайно высоко. Вот что говорит по этому поводу Лаплас: «В своей бинарной арифметике Лейбниц видел прообраз творения. Ему представлялось, что единица представляет божественное начало, а нуль — небытие, и что Высшее Существо создает все сущее из небытия точно таким же образом, как единица и нуль в его системе выражают все числа».

Упражнение. Исследуйте в общем виде вопрос о представлении чисел в системе с основанием a. Чтобы называть числа в этой системе, нужны наименования для однозначных чисел $0, 1, \ldots, a-1$ и для различных степеней a: a, a^2, a^3, ... Сколько именно числительных потребуется, чтобы назвать все числа до одной тысячи в системах с основанием $a = 2, 3, 4, 5, \ldots, 15$? Каково должно быть основание a, чтобы число этих имен числительных было наименьшим? (*Примеры*: если $a = 10$, то нужно десять числительных для однозначных чисел. Затем еще три числительных, обозначающих 10, 100 и 1000, всего — 13. При $a = 20$ нужно двадцать числительных для однозначных чисел и еще числительные для 20 и 400; всего — 22. При $a = 100$ понадобится 101 числительное.)

§ 2. Бесконечность системы натуральных чисел. Математическая индукция

1. Принцип математической индукции. Последовательность натуральных чисел 1, 2, 3, 4, ... не имеет конца: действительно, как только достигается некоторое число n, вслед за ним сейчас же можно написать ближайшее к нему натуральное число $n+1$. Желая как-нибудь назвать эти свойства последовательности натуральных чисел, мы говорим, что этих чисел существует *бесконечное множество.* Последовательность натуральных чисел представляет простейший и самый естественный пример бесконечного (в математическом смысле), играющего господствующую роль в современной математике. Не раз в этой книге нам придется иметь дело с совокупностями, содержащими бесконечное множество объектов; такова, например, совокупность всех точек на прямой линии или совокупность всех треугольников на плоскости. Но бесконечная последовательность натуральных чисел безусловно представляет простейший пример бесконечной совокупности.

Последовательный, шаг за шагом, переход от n к $n+1$, порождающий бесконечную последовательность натуральных чисел, вместе с тем лежит в основе одного из важнейших и типичных для математики рассуждений, именно принципа математической индукции. «Эмпирическая индукция», применяемая в естественных науках, исходит из частного ряда наблюдений некоторого явления и приходит к констатации общего закона, которому подчиняется явление в его различных формах. Степень уверенности, с которой закон таким образом устанавливается, зависит от числа отдельных наблюдений и выводимых из них заключений. Часто подобного рода индуктивные рассуждения бывают вполне убедительными; утверждение, что солнце взойдет завтра с востока, столь несомненно, насколько это вообще возможно; и все же характер констатации в данном случае совсем иной, чем в случае теоремы, доказываемой на основе строгого логического, т. е. математического, рассуждения.

Что касается *математической индукции*, то она применяется иным, отличным способом с целью установления истинности математической теоремы в бесконечной последовательности случаев (первого, второго, третьего и так далее — без всякого исключения). Обозначим через A некоторое утверждение, относящееся к произвольному натуральному числу n. Пусть A будет хотя бы такое утверждение: «Сумма углов в выпуклом многоугольнике с $n+2$ сторонами равна $180° \cdot n$». Или еще: обозначим через A' утверждение: «проводя n прямых на плоскости, нельзя разбить ее больше чем на 2^n частей». Чтобы доказать подобного рода теорему для *произвольного* значения n, недостаточно доказать ее отдельно для первых 10, или 100, или даже 1000 значений n. Это как раз соответствовало бы принципу эмпирической индукции. Вместо того нам приходится воспользоваться строго математическим и отнюдь не эмпирическим рассуждением; мы уясним себе его характер на частных примерах доказательства предложений, которые мы обозначили через A и A'. Остановимся на предложении A. Если $n = 1$, то речь идет о треугольнике, и мы знаем из элементарной геометрии, что сумма углов такового равна $180° \cdot 1$. В случае четырехугольника, ($n = 2$) мы проводим диагональ, разделяющую четырехугольник на два треугольника, и тогда сейчас же становится ясно, что сумма углов четырехугольника равна сумме углов в двух треугольниках, именно равна $180° + 180° = 180° \cdot 2$. Обращаясь к случаю пятиугольника ($n = 3$), мы разбиваем его таким же образом на четырехугольник и треугольник. Так как первый из названных многоугольников по доказанному имеет сумму углов $180° \cdot 2$, а второй — $180° \cdot 1$, то всего в случае пятиугольника мы получаем сумму углов $180° \cdot 3$. И теперь нам уже становится ясно, что рассуждение может быть продолжено совершенно таким же образом неограниченно. Мы докажем теорему для случая $n = 4$, затем для случая $n = 5$, и т. д. Как и раньше,

каждое следующее заключение неизбежно вытекает из предыдущего, и теорема A оказывается установленной при произвольном значении n.

Так же обстоит дело и с предложением A'. При $n = 1$ оно, очевидно, справедливо, так как всякая прямая делит плоскость на 2 части. Проведем вторую прямую. Каждая из двух прежних частей разобьется в свою очередь на две части — при условии, что вторая прямая непараллельна первой. Но, как бы то ни было, в случае $n = 2$ всего окажется не более $4 = 2^2$ частей. Добавим еще третью прямую. Каждая из уже имеющихся частей или будет разбита на две части, или останется нетронутой. Таким образом, число вновь полученных частей не превысит $2^2 \cdot 2 = 2^3$. Считая это установленным, мы точно так же перейдем к следующему случаю и т. д. — без конца.

Сущность предыдущего рассуждения заключается в том, что, желая установить справедливость некоторой общей теоремы A при любых значениях n, мы доказываем эту теорему последовательно для бесконечного ряда специальных случаев A_1, A_2, ... Возможность этого рассуждения покоится на двух предпосылках: а) имеется общий метод доказательства того, что *если* справедливо утверждение A_r, то следующее по порядку утверждение A_{r+1} *также* справедливо; б) *известно*, что первое утверждение A_1 справедливо. В том, что эти два условия достаточны для того, чтобы справедливость *всех* утверждений A_1, A_2, A_3, ... была установлена, заключается некоторый логический принцип, имеющий в математике столь же фундаментальное значение, как и классические правила аристотелевой логики.

Сформулируем этот принцип следующим образом. Предположим, что требуется установить справедливость бесконечной последовательности математических предложений

$$A_1, \ A_2, \ A_3, \ \ldots,$$

которые все, совместно взятые, образуют некоторое общее предложение A. *Допустим, что а) проведено математическое рассуждение, показывающее, что если верно A_r, то верно и A_{r+1}, каково бы ни было натуральное число r, и б) установлено, что A_1 верно. Тогда все предложения нашей последовательности верны и, следовательно, предложение A доказано.*

Мы примем принцип индукции без колебаний (так же как мы принимаем все правила обыкновенной логики) и будем его рассматривать как основной принцип, на котором строится математическое доказательство. В самом деле, мы можем установить справедливость каждого утверждения A_n, исходя из допущения б) о том, что A_1 справедливо, и, многократно пользуясь допущением а), последовательно установим справедливость утверждений A_2, A_3, A_4, и т. д., пока не достигнем утверждения A_n. Принцип математической индукции вытекает, таким

образом, из того факта, что за каждым натуральным числом r следует (непосредственно) другое натуральное число $r+1$ и что, отправляясь от натурального числа 1, можно после конечного числа таких переходов достигнуть любого натурального числа n.

Часто принцип математической индукции применяют без явного о том упоминания или же просто он скрывается за формулой «и так далее». Такая скрытая форма применения принципа индукции в особенности свойственна преподаванию элементарной математики. Но при доказательстве иных, более тонких и более глубоких теорем этим принципом неизбежно приходится пользоваться явно. Мы приведем далее некоторое число относящихся сюда простых и все же не совсем тривиальных примеров.

2. Арифметическая прогрессия. *Каково бы ни было значение* n, *сумма* $1+2+3+\ldots+n$ *первых* n *натуральных чисел равна* $\dfrac{n(n+1)}{2}$. Чтобы доказать эту теорему по принципу математической индукции, мы должны для произвольного значения n установить справедливость соотношения A_n:

$$1+2+3+\ldots+n=\frac{n(n+1)}{2}. \tag{1}$$

а) *Если* r — некоторое натуральное число и *если* известно, что утверждение A_r справедливо, т. е. если известно, что

$$1+2+3+\ldots+r=\frac{r(r+1)}{2},$$

то, прибавляя к обеим частям последнего равенства по $r+1$, мы получаем:

$$1+2+3+\ldots+r+(r+1)=\frac{r(r+1)}{2}+(r+1)= \\ =\frac{r(r+1)+2(r+1)}{2}=\frac{(r+1)(r+2)}{2},$$

а это как раз и есть утверждение A_{r+1}.

б) Утверждение A_1, очевидно, справедливо, так как $1=\dfrac{1\cdot 2}{2}$. Итак, по принципу математической индукции утверждение A_n справедливо при любом n, что и требовалось доказать.

Обыкновенно эту теорему доказывают иным способом. Пишут сумму $1+2+3+\ldots+n$ в двух видах:

$$S_n=1+2+\ldots+(n-1)+n$$

и

$$S_n=n+(n-1)+\ldots+2+1.$$

Складывая, мы видим, что числа, стоящие на одной вертикали, вместе составляют $n+1$, и так как вертикалей всего имеется n, то отсюда

следует, что

$$2S_n = n(n+1),$$

и остается еще разделить на 2.

Из формулы (1) сразу же вытекает общая формула для суммы $(n+1)$ первых членов любой арифметической прогрессии:

$$P_n = a + (a+d) + (a+2d) + \ldots + (a+nd) = \frac{(n+1)(2a+nd)}{2}. \qquad (2)$$

В самом деле,

$$P_n = (n+1)a + (1+2+\ldots+n)d = (n+1)a + \frac{n(n+1)d}{2} = \\ = \frac{2(n+1)a + n(n+1)d}{2} = \frac{(n+1)(2a+nd)}{2}.$$

В случае, когда $a=0$, $d=1$, последнее соотношение превращается в соотношение (1).

3. Геометрическая прогрессия. Таким же образом можно рассуждать и по поводу геометрической прогрессии (в общем виде). Мы покажем, что, каково бы ни было n,

$$G_n = a + aq + aq^2 + \ldots + aq^n = a\,\frac{1-q^{n+1}}{1-q}. \qquad (3)$$

(Мы предполагаем, что $q \neq 1$: иначе правая часть (3) лишена смысла.)

Наше утверждение, несомненно, справедливо при $n=1$, так как в этом случае

$$G_1 = a + aq = \frac{a(1-q^2)}{1-q} = \frac{a(1+q)(1-q)}{1-q} = a(1+q).$$

И *если* мы допустим, что

$$G_r = a + aq + \ldots + aq^r = a\,\frac{1-q^{r+1}}{1-q},$$

то, как следствие, отсюда немедленно вытекает:

$$G_{r+1} = (a + aq + \ldots + aq^r) + aq^{r+1} = G_r + aq^{r+1} = \\ = a\,\frac{1-q^{r+1}}{1-q} + aq^{r+1} = a\,\frac{(1-q^{r+1}) + q^{r+1}(1-q)}{1-q} = \\ = a\,\frac{1-q^{r+1}+q^{r+1}-q^{r+2}}{1-q} = a\,\frac{1-q^{r+2}}{1-q}.$$

Но это как раз и есть утверждение (3) при $n = r+1$. Доказательство закончено.

В элементарных учебниках дается другое доказательство. Положим

$$G_n = a + aq + \ldots + aq^n.$$

Умножая обе части на q, получим

$$qG_n = aq + aq^2 + \ldots + aq^{n+1}.$$

Вычитая затем последнее равенство из предпоследнего, получаем далее

$$G_n - qG_n = a - aq^{n+1},$$
$$(1-q)G_n = a(1-q^{n+1}),$$
$$G_n = a\,\frac{1-q^{n+1}}{1-q}.$$

4. Сумма n первых квадратов. Следующее интересное применение принципа математической индукции относится к сумме n первых квадратов. Путем проб мы устанавливаем (по крайней мере для нескольких небольших значений n), что

$$1^2 + 2^2 + 3^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}, \tag{4}$$

после чего естественно высказать в виде догадки утверждение, что эта замечательная формула справедлива при *всех целых положительных значениях* n. Чтобы *доказать* это, воспользуемся опять методом математической индукции. Заметим прежде всего, что *если* утверждение A_n, которое заключается как раз в соотношении (4), справедливо при $n = r$, так что

$$1^2 + 2^2 + 3^2 + \ldots + r^2 = \frac{r(r+1)(2r+1)}{6},$$

то, прибавляя к обеим частям по $(r+1)^2$, мы получаем

$$1^2 + 2^2 + \ldots + r^2 + (r+1)^2 = \frac{r(r+1)(2r+1)}{6} + (r+1)^2 =$$
$$= \frac{r(r+1)(2r+1) + 6(r+1)^2}{6} = \frac{(r+1)[r(2r+1) + 6(r+1)]}{6} =$$
$$= \frac{(r+1)(2r^2+7r+6)}{6} = \frac{(r+1)(r+2)(2r+3)}{6},$$

а это и есть утверждение A_{r+1}, так как оно получается из соотношения (4) посредством подстановки $r+1$ вместо n. Чтобы закончить доказательство, достаточно обратить внимание на то, что утверждение A_1, которое сводится к равенству

$$1^2 = \frac{1(1+1)(2+1)}{6},$$

справедливо. Итак, соотношение (4) верно при всех значениях n.

Подобного же рода формулы можно написать для сумм третьих и четвертых степеней, вообще для сумм вида $1^k + 2^k + 3^k + \ldots + n^k$, где k — произвольное целое положительное число. В качестве упражнения читатель может доказать с помощью математической индукции формулу

$$1^3 + 2^3 + 3^3 + \ldots + n^3 = \left(\frac{n(n+1)}{2}\right)^2. \tag{5}$$

Необходимо заметить в заключение, что, хотя принципа математической индукции совершенно достаточно для того, чтобы *доказать* формулу (5) — раз она уже написана, однако доказательство не дает решительно никаких указаний, как прийти к самой этой формуле: почему именно нужно догадываться, что сумма n первых кубов равна выражению $\left(\frac{n(n+1)}{2}\right)^2$, а не какому-нибудь иному в таком же роде, например, $\left(\frac{n(n+1)}{3}\right)^2$ или $\frac{19n^2-41n+24}{3}$, и т. д. Выбор велик! Тот факт, что доказательство теоремы заключается в применении таких-то простых логических правил, не оказывает ни малейшего влияния на творческое начало в математике, роль которого — делать выбор из бесконечного множества возникающих возможностей. Вопрос о том, как возникает *гипотеза* (5), принадлежит к той области, в которой нет никаких общих правил; здесь делают свое дело эксперимент, аналогия, конструктивная индукция. Раз только правильная гипотеза сформулирована, принципа математической индукции часто бывает достаточно, чтобы теорема была доказана. Но так как само такое доказательство никак не указывает пути к открытию, то его лучше было бы называть *проверкой*.

***5. Одно важное неравенство.** В следующей главе нам понадобится неравенство

$$(1+p)^n \geqslant 1+np, \tag{6}$$

имеющее место при всяком p, удовлетворяющем условию $p > -1$, и при любом целом положительном значении n. (Ради общности мы предвосхищаем здесь применение отрицательных и нецелых чисел, предполагая, что p может быть любым числом, большим чем -1. Доказательство неравенства одно и то же, независимо от того, каково число p.) Мы воспользуемся и на этот раз математической индукцией.

а) *Если* верно, что $(1+p)^r \geqslant 1+rp$, то, умножая обе части неравенства на положительное число $1+p$, мы получаем:

$$(1+p)^{r+1} \geqslant 1+rp+p+rp^2.$$

Отбрасывая вовсе положительный член rp^2, мы только усилим это неравенство; итак,

$$(1+p)^{r+1} \geqslant 1+(r+1)p.$$

Полученный результат показывает, что неравенство (6) имеет место и при $n = r+1$.

б) Совершенно очевидно, что $(1+p)^1 \geqslant 1+p$. Таким образом, доказательство закончено.

Ограничение, заключающееся в условии $p > -1$, существенно. Если $p < -1$, то $1+p$ отрицательно, и рассуждение а) отпадает, так как при умножении обеих частей неравенства на отрицательное число знак

неравенства должен измениться. (Например, умножая обе части неравенства $3 > 2$ на -1, мы получили бы $-3 > -2$, а это неверно.)

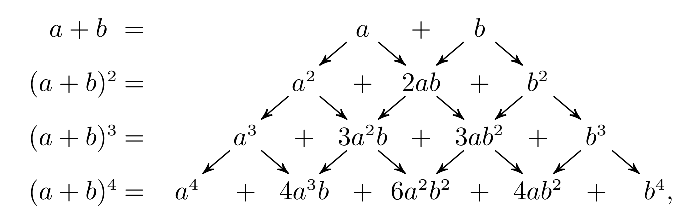
***6. Биномиальная теорема.** Часто бывает нужно написать в раскрытом виде выражение для n-й степени бинома $(a+b)^n$. Непосредственное вычисление показывает, что

при $n=1$ $\quad (a+b)^1 = a+b,$

при $n=2$ $\quad (a+b)^2 = (a+b)(a+b) =$
$$= a(a+b) + b(a+b) = a^2 + 2ab + b^2,$$

при $n=3$ $\quad (a+b)^3 = (a+b)(a+b)^2 =$
$$= a(a^2+2ab+b^2) + b(a^2+2ab+b^2) = a^3 + 3a^2b + 3ab^2 + b^3,$$

и так далее. Но какой общий закон скрывается за словами «и так далее»? Проанализируем процесс вычисления $(a+b)^2$. Так как $(a+b)^2 = (a+b)(a+b)$, то мы получили выражение для $(a+b)^2$, умножая каждый член выражения $a+b$ на a, затем на b и складывая то, что получилось. Ту же процедуру пришлось применить при вычислении $(a+b)^3 = (a+b)(a+b)^2$. Так же точно вычисляются $(a+b)^4$, $(a+b)^5$ и так далее до бесконечности. Выражение для $(a+b)^n$ мы получим, умножая выражение $(a+b)^{n-1}$ сначала на a, потом на b, затем складывая то, что получится. Это приводит к следующей диаграмме:



позволяющей сразу разобраться в общем законе составления коэффициентов в разложении $(a+b)^n$. Мы строим треугольную схему из натуральных чисел, начиная с коэффициентов 1, 1 двучлена $a+b$ таким образом, что каждое число в треугольнике является суммой двух чисел, стоящих над ним в предыдущей строке (слева и справа). Такая схема известна под названием *треугольника Паскаля.*

1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
. . . .

Коэффициенты в разложении $(a + b)^n$ *по убывающим степеням* a *и возрастающим степеням* b *стоят в* n*-й строке этой схемы.*

Так, например,

$$(a+b)^7 = a^7 + 7a^6b + 21a^5b^2 + 35a^4b^3 + 35a^3b^4 + 21a^2b^5 + 7ab^6 + b^7.$$

С помощью очень сжатых обозначений, использующих нижние и верхние значки (индексы), запишем числа, стоящие в n-й строке треугольника Паскаля, следующим образом:

$$C_0^n = 1,\ \ C_1^n,\ \ C_2^n,\ \ C_3^n,\ \ \ldots,\ \ C_{n-1}^n,\ \ C_n^n = 1.$$

Тогда общей формуле для разложения $(a + b)^n$ можно придать вид

$$(a+b)^n = a^n + C_1^n a^{n-1}b + C_2^n a^{n-2}b_2 + \ldots + C_{n-1}^n ab^{n-1} + b^n. \qquad (7)$$

Согласно закону, лежащему в основе построения треугольника Паскаля, мы имеем соотношение

$$C_i^n = C_{i-1}^{n-1} + C_i^{n-1}. \qquad (8)$$

В качестве упражнения читатель, имеющий уже некоторый опыт в применении математической индукции, может воспользоваться этим принципом (и очевидными равенствами $C_0^1 = C_1^1 = 1$) для доказательства общей формулы

$$C_i^n = \frac{n(n-1)(n-2)\ldots(n-i+1)}{1\cdot 2\cdot 3\cdot\ldots\cdot i} = \frac{n!}{i!\,(n-i)!}. \qquad (9)$$

(При любом целом положительном значении n символ $n!$ (читается «n-факториал») обозначает произведение n первых натуральных чисел: $n! = 1\cdot 2\cdot 4\cdot\ldots\cdot n$. Удобно положить по определению $0! = 1$, чтобы формула (9) оправдывалась также и при i, равном 0 или n.)

Выводу этой раскрытой формулы для коэффициентов биномиального разложения иногда дается наименование *биномиальной теоремы* (см. также стр. 500).

Упражнения. Докажите с помощью метода математической индукции следующие равенства:

1) $\dfrac{1}{1\cdot 2} + \dfrac{1}{2\cdot 3} + \ldots + \dfrac{1}{n(n+1)} = \dfrac{n}{n+1}$.

2) $\dfrac{1}{2} + \dfrac{2}{2^2} + \dfrac{3}{2^3} + \ldots + \dfrac{n}{2^n} = 2 - \dfrac{n+2}{2^n}$.

*3) $1 + 2q + 3q^2 + \ldots + nq^{n-1} = \dfrac{1-(n+1)q^n + nq^{n+1}}{(1-q)^2}$.

*4) $(1+q)(1+q^2)(1+q^4)\ldots(1+q^{2^n}) = \dfrac{1-q^{2^{n+1}}}{1-q}$.

Найдите сумму следующих геометрических прогрессий:

5) $\dfrac{1}{1+x^2} + \dfrac{1}{(1+x^2)^2} + \ldots + \dfrac{1}{(1+x^2)^n}$.

6) $1 + \dfrac{x}{1+x^2} + \dfrac{x^2}{(1+x^2)^2} + \ldots + \dfrac{x^n}{(1+x^2)^n}$.

7) $\frac{x^2-y^2}{x^2+y^2}+\left(\frac{x^2-y^2}{x^2+y^2}\right)^2+\ldots+\left(\frac{x^2-y^2}{x^2+y^2}\right)^n$.

Пользуясь формулами (4) и (5), докажите равенства:

*8) $1^2+3^2+\ldots+(2n+1)^2=\frac{(n+1)(2n+1)(2n+3)}{3}$.

*9) $1^3+3^3+\ldots+(2n+1)^3=(n+1)^2\,(2n^2+4n+1)$.

10) То же самое докажите непосредственно методом математической индукции.

7. Дальнейшие замечания по поводу метода математической индукции. Принцип математической индукции может быть слегка обобщен следующим образом: *если имеется последовательность утверждений A_s, A_{s+1}, A_{s+2}, ..., где s — некоторое положительное число, и если*

а) *при всяком значении $r \geqslant s$ справедливость A_{r+1} следует из справедливости A_r, и*

б) *известно, что A_s справедливо,*

то все утверждения A_s, A_{s+1}, A_{s+2}, ... справедливы. Иначе говоря, A_n «справедливо при любом $n \geqslant s$». То же самое рассуждение, которое привело нас к обыкновенному принципу математической индукции, пригодно и в данном случае, только последовательность 1, 2, 3, ... заменяется на этот раз подобной ей последовательностью s, $s+1$, $s+2$, $s+3$, ...

Пользуясь принципом индукции в этой форме, мы можем усилить неравенство (6) на стр. 34, исключая возможность знака равенства. Именно: *каково бы ни было $p \neq 0$ и > -1 и каково бы ни было целое число $n \geqslant 2$, имеет место неравенство*

$$(1+p)^n > 1+np. \tag{10}$$

Доказательство предоставляется читателю.

С принципом математической индукции тесно связан «принцип наименьшего целого числа», утверждающий, что *во всяком непустом множестве C натуральных чисел имеется наименьшее число.* Множество C может быть конечным, как, например, множество 1, 2, 3, 4, 5, или бесконечным, как, например множество всех четных чисел 2, 4, 6, 8, 10, ... Множество называется *пустым*, если оно не содержит ни одного элемента; примером пустого множества может служить множество всех кругов, одновременно являющихся прямыми линиями, или множество натуральных чисел n, удовлетворяющих соотношению $n > n$. По понятным причинам мы оговариваем в формулировке «принципа наименьшего целого числа», что пустые множества исключаются. Всякое непустое множество C целых чисел непременно содержит хоть одно число, например n, и тогда наименьшее из чисел 1, 2, 3, ..., n, принадлежащее множеству C, есть наименьшее целое число множества.

Чтобы уяснить значение этого принципа, следует указать на то, что он, вообще говоря, неверен для множеств, состоящих из нецелых чисел; например, множество положительных дробных чисел 1, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, ... не содержит наименьшего числа.

С чисто логической точки зрения небезынтересно отметить то обстоятельство, что с помощью принципа наименьшего целого числа принцип математической индукции *доказывается* как теорема. В самом деле, пусть имеется

последовательность таких утверждений A_1, A_2, A_3, ..., что

а) при любом r справедливость A_{r+1} вытекает из справедливости A_r,

б) известно, что A_1 справедливо.

Мы докажем, что предположение о том, что хоть одно из утверждений A несправедливо, придется отбросить. Действительно, если бы хоть одно из утверждений A было неверным, то множество всех натуральных чисел n, для которых A_n неверно, не было бы пустым. Тогда согласно принципу наименьшего целого числа оно содержало бы наименьшее число p, которое вследствие б) должно было бы быть больше чем 1. Но тогда A_p было бы неверно, а A_{p-1} верно. Это противоречит условию а).

Подчеркнем еще раз, что принцип математической индукции резко отличается от эмпирической индукции, свойственной естественным наукам. Подтверждение общего закона на конечном числе случаев (как бы это число ни было велико) никоим образом не представляет собой доказательства в математическом смысле, даже если неизвестно ни одного исключения. При таких обстоятельствах рассматриваемое утверждение, или «закон», есть не что иное, как вполне разумная *гипотеза*, которую могут видоизменить результаты будущих экспериментов. В математике «закон» может считаться доказанным только тогда, когда он выведен как неизбежное логическое следствие из предпосылок, признаваемых справедливыми. Существует немало примеров математических утверждений, которые были проверены и оправдывались во всех до настоящего времени рассмотренных частных случаях, но для которых еще не было найдено общего доказательства (см. пример на стр. 49). Можно *подозревать*, что теорема справедлива во всей общности, если она подтверждается на большом числе примеров, и тогда есть основание пытаться *доказать* ее с помощью математической индукции. Если попытка удается, то тем самым справедливость теоремы устанавливается; в противном случае вопрос о том, верна или неверна теорема, остается открытым, и она может быть доказана или опровергнута когда-нибудь в будущем уже иными методами.

Применяя принцип математической индукции, необходимо всегда тщательно следить за тем, чтобы условия а) и б) были действительно выполнены. Иначе можно иной раз прийти и к абсурду. Мы предлагаем читателю разобраться в следующем софизме. Мы «докажем» сейчас, что *любые два целых положительных числа равны между собой*; например, $5 = 10$. Начнем с определения. Если a и b — два неравных между собой целых положительных числа, то через $\max(a, b)$ будем обозначать a или b, смотря по тому, какое из чисел больше: a или b; если же $a = b$, то положим $\max(a, b) = a = b$. Так, $\max(3, 5) = \max(5, 3) = 5$, тогда как $\max(4, 4) = 4$. Далее, через A_n обозначим следующее утверждение: «Если a и b — два таких целых положительных числа, что $\max(a, b) = n$, то $a = b$».

а) Предположим, что A_r верно. Пусть a и b — два таких целых положи-

тельных числа, что $\max(a, b) = r + 1$. Рассмотрим числа

$$\alpha = a - 1, \qquad \beta = b - 1;$$

тогда $\max(\alpha, \beta) = r$. В таком случае $\alpha = \beta$, так как A_r верно. Но отсюда следует, что $a = b$; значит, верно и A_{r+1}.

б) A_1, очевидно, верно, так как если $\max(a, b) = 1$, то a и b (по предположению — целые положительные числа) должны быть каждое в отдельности равны 1.

Итак, по принципу математической индукции A_n верно при любом n.

Пусть теперь a и b — два произвольных целых положительных числа; положим $\max(a, b) = r$. Было показано, что A_n верно при любом n; значит, в частности, верно A_r. Следовательно, $a = b$.

ДОПОЛНЕНИЕ К ГЛАВЕ I

Теория чисел

Введение

Мистические и суеверные представления, связывавшиеся первоначально с целыми числами, мало-помалу изгладились, но среди математиков интерес к числам не ослабевал никогда. Евклид (около 300 г. до нашей эры), громкая слава которого объясняется той частью его «Начал», которая посвящена основам геометрии (изучаемым в школе), по-видимому, сделал оригинальные открытия в области теории чисел, тогда как его геометрия в значительной степени представляет собой компиляцию ранее полученных результатов. Диофант Александрийский (около 275 г. нашей эры), один из первых алгебраистов, оставил также след в теории чисел. Пьер Ферма (1601–1665), живший в Тулузе, по специальности юрист, вместе с тем замечательнейший математик своей эпохи, положил начало современным теоретико-числовым изысканиям. Эйлер (1707–1783), наиболее изумительный из математиков в смысле богатства продукции, в своих исследованиях весьма часто углублялся в область теории чисел. Сюда же следует прибавить ряд иных имен, знаменитых в анналах математики: Лежандр, Дирихле, Риман. Гаусс (1777–1855), виднейший из математиков более близкой к нам эпохи, в равной степени отдававший себя различным отраслям математики, следующими словами определил свое отношение к теории чисел: «Математика — царица наук, теория чисел — царица математики».

§ 1. Простые числа

1. Основные факты. Многие утверждения в области теории чисел, как и математики вообще, относятся не к отдельным объектам, скажем, к числу 5 или числу 32, а к целому классу объектов, имеющих какое-то общее свойство; примерами могут служить класс всех четных чисел

$$2,\ 4,\ 6,\ 8,\ \ldots,$$

или класс чисел, делящихся на 3,

$$3,\ 6,\ 9,\ 12,\ \ldots,$$

или класс квадратов целых чисел

$$1,\ 4,\ 9,\ 16,\ \ldots$$

и так далее.

В теории чисел особенно важную роль играет класс всех *простых чисел*. Очень многие числа могут быть разложены на меньшие множители: $10 = 2 \cdot 5$, $111 = 3 \cdot 37$, $144 = 3 \cdot 3 \cdot 2 \cdot 2 \cdot 2 \cdot 2$, и т. п. Числа, которые таким образом *не* разлагаются, носят название простых. Точнее, *простым называется такое целое число* p, *большее единицы, которое не имеет иных множителей, кроме единицы и самого себя.* (Число *a есть множитель*, или *делитель*, числа b или *делит* число b, если существует такое *целое* число c, что $b = ac$.) Числа 2, 3, 5, 7, 11, 13, 17, — простые, тогда как, например, число 12 не является простым, так как $12 = 3 \cdot 4$. Значение класса простых чисел заключается в том, что *каждое* число может быть представлено как *произведение простых*: если данное число не простое, то его можно последовательно разлагать на множители до тех пор, пока все множители не окажутся простыми; так, например, $360 = 3 \cdot 120 = 3 \cdot 30 \cdot 4 = 3 \cdot 3 \cdot 10 \cdot 2 \cdot 2 = 3 \cdot 3 \cdot 5 \cdot 2 \cdot 2 \cdot 2 = 2^3 \cdot 3^2 \cdot 5$. Число, отличное от 0 и 1 и не являющееся простым, называется *составным.*

Один из первых вопросов, возникающих по поводу класса простых чисел, заключается в том, существует ли только конечное число различных простых чисел или же класс простых чисел содержит бесконечное число членов подобно классу всех целых чисел, частью которого он является. Ответ таков: *простых чисел существует бесконечное множество.*

Данное Евклидом доказательство существования бесконечного множества простых чисел представляет собой типичный образец математического рассуждения. В основе его лежит «косвенный метод» (доказательство от противного, приведение к абсурду). Сделаем попытку допустить, что рассматриваемое предложение неверно. Это означало бы, что существует лишь конечное число простых чисел, хотя, может быть, и очень много — скажем, около миллиарда; тогда допустим, что это число, представленное в «общей» или «неопределенной» форме, будет n.

Пользуясь знаками, мы можем обозначить все простые числа через p_1, $p_2, \ldots, p_n$. Всякое иное число тем самым составное и должно делиться по меньшей мере на одно из простых чисел $p_1, p_2, \ldots, p_n$. А теперь мы придем к противоречию, а именно, построим число A, которое будет отлично от каждого из чисел $p_1, p_2, \ldots, p_n$, так как будет больше их всех и тем не менее не будет делиться ни на одно из них. Вот это число:

$$A = p_1 p_2 \ldots p_n + 1.$$

Как видно, оно равно единице плюс произведение тех чисел, которые образуют совокупность всех простых чисел. Число A больше, чем любое из чисел p, и потому должно быть составным. Но при делении на p_1, на p_2 и т. д. A дает всякий раз остаток 1, таким образом, A не делится ни на одно из чисел p. Сделанное нами допущение, что существует лишь конечное число простых чисел, приводит, таким образом, к противоречию, так что приходится заключить, что это допущение ошибочно, а следовательно, истинным может быть только противоположное ему. Итак, теорема доказана.

Хотя это доказательство и «косвенного» характера, все же небольшое его видоизменение приводит, по крайней мере теоретически, к методу построения бесконечной последовательности простых чисел. Предположим, что, исходя из некоторого простого числа, скажем $p_1 = 2$, мы уже нашли n простых чисел p_1, $p_2, \ldots, p_n$; заметим, далее, что число $p_1 p_2 \ldots p_n + 1$ или простое, или содержит множителем простое число, отличное от тех, которые уже найдены. Так как такой множитель всегда может быть найден (хотя бы непосредственными пробами), то в обоих названных случаях мы в итоге получаем новое простое число p_{n+1}; продолжая таким же образом дальше, убеждаемся, что последовательность простых чисел, которые мы действительно можем построить, не имеет конца.

Упражнение. Выполните намеченное построение, начиная с простых чисел $p_1 = 2$, $p_2 = 3$; найдите еще пять простых чисел.

Если какое-нибудь число представлено в виде произведения простых множителей, то эти множители можно, конечно, располагать в каком угодно порядке. Занимаясь разложением чисел на простые множители, мы очень скоро приходим к заключению, что с точностью до порядка сомножителей разложение любого числа N на простые множители обладает свойством единственности: *каждое натуральное число N, большее единицы, может быть разложено на простые множители только одним способом.* Это утверждение кажется на первый взгляд таким очевидным, что неспециалист склонен обыкновенно отвергать необходимость его доказательства. Все же рассматриваемое предложение отнюдь не тривиально; с другой стороны, его доказательство, хотя и совсем элементарного содержания, требует более или менее тонких рассуждений. Классическое доказательство этой «основной теоремы арифмети-

ки», данное Евклидом, базируется на методе («алгоритме») нахождения общего наибольшего делителя двух чисел. Этот метод будет нами рассмотрен на стр. 61. Здесь же мы приведем доказательство менее почтенной давности, которое несколько короче, чем доказательство Евклида, и вместе с тем выглядит несколько более «софистическим». Оно также является типичным образцом «косвенного» рассуждения. Мы допустим, что существует такое число, которое может быть разложено на простые множители двумя существенно различными способами, и это допущение приведет нас к противоречию. Возникновение противоречия будет свидетельствовать о том, что гипотеза о существовании числа, допускающего два существенно различных разложения на простые множители, несостоятельна; и отсюда мы заключим, что разложение чисел на простые множители обладает свойством единственности.

* Если существует хоть одно число, допускающее два существенно различных разложения на простые множители, то существует непременно и *наименьшее* число, обладающее таким свойством (см. стр. 37),

$$m = p_1 p_2 \ldots p_r = q_1 q_2 \ldots q_s, \tag{1}$$

где через p и q обозначены простые числа. Меняя, если потребуется, порядок этих множителей, мы можем допустить, что

$$p_1 \leqslant p_2 \leqslant \ldots \leqslant p_r, \qquad q_1 \leqslant q_2 \leqslant \ldots \leqslant q_s.$$

Заметим, что p_1 отлично от q_1: иначе, деля равенство (1) на общий простой множитель, мы получили бы два существенно различных разложения на простые множители числа, которое было бы меньше, чем m, и это противоречило бы предложению о том, что m — наименьшее число, обладающее таким свойством. Следовательно, одно из двух: или $p_1 < q_1$, или $q_1 < p_1$. Пусть $p_1 < q_1$. (Если бы оказалось $q_1 < p_1$, то в дальнейшем рассуждении достаточно было бы поменять местами буквы p и q.) Рассмотрим целое число

$$m' = m - (p_1 q_2 q_3 \ldots q_s). \tag{2}$$

Подставляя вместо m два его выражения, взятые из равенства (1), мы можем представить число m' в любом из двух видов:

$$m' = (p_1 p_2 \ldots p_r) - (p_1 q_2 q_3 \ldots q_s) = p_1 (p_2 \ldots p_r - q_2 q_3 \ldots q_s), \tag{3}$$

$$m' = (q_1 q_2 \ldots q_s) - (p_1 q_2 q_3 \ldots q_s) = (q_1 - p_1) q_2 q_3 \ldots q_s. \tag{4}$$

Из равенства (4) следует, что m' — положительное число, так как $p_1 < q_1$; из равенства (2) следует, с другой стороны, что m' меньше чем m. Но раз так, то разложение m' на множители должно быть *единственным* (с точностью до порядка сомножителей). Из равенства (3) далее видно, что p_1 входит множителем в m'; значит, из равенства (4) можно в таком случае заключить, что p_1 входит множителем или в $q_1 - p_1$, или

в $q_2q_3\ldots q_s$. (Это вытекает из единственности разложения m' на простые множители; см. рассуждение в следующем абзаце.) Но последнее невозможно, так как все q больше чем p_1. Поэтому p_1 должно входить множителем в $q_1 - p_1$, т. е. $q_1 - p_1$ должно делиться на p_1. Другими словами, существует такое целое число h, что

$$q_1 - p_1 = p_1 \cdot h, \quad \text{или} \quad q_1 = p_1(h+1).$$

Но это значит, что q_1 делится на p_1, чего, однако, быть не может, так как, по предположению, q_1 — число простое. Противоречие, к которому мы пришли, показывает несостоятельность первоначально сделанного допущения, чем и заканчивается доказательство основной теоремы арифметики.

Вот одно важное следствие основной теоремы. *Если простое число p входит множителем в произведение ab, то оно непременно входит множителем или в a, или в b.* В самом деле, если бы p не входило множителем ни в a, ни в b, то, перемножая разложения на простые множители чисел a и b, мы получили бы разложение на простые множители числа ab, *не содержащее множителя p*. С другой стороны, так как предполагается, что p входит множителем в произведение ab, то это значит, что существует такое целое число t, что

$$ab = pt.$$

Поэтому, перемножая p и разложение на простые множители числа t, мы получим разложение на простые множители числа ab, *содержащее множитель p*. Таким образом, приходится признать, что существует два различных разложения числа ab на простые множители, а это противоречит основной теореме.

Примеры. Если установлено, что 2652 делится на 13 и что $2652 = 6 \cdot 442$, то отсюда можно сделать заключение, что 442 делится на 13. С другой стороны, 240 делится на 6 и притом $240 = 15 \cdot 16$, но ни 15, ни 16 не делятся на 6. Этот пример показывает, что условие основной теоремы относительно того, что число p — простое, является существенным.

Упражнение. Чтобы найти все делители числа a, достаточно разложить a в произведение

$$a = p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_r^{\alpha_r},$$

где все множители p — простые и различные, причем каждый из них возводится в некоторую степень. Все делители числа a имеют вид

$$b = p_1^{\beta_1} p_2^{\beta_2} \ldots p_r^{\beta_r},$$

где показатели β — произвольные целые числа, подчиненные условиям

$$0 \leqslant \beta_1 \leqslant \alpha_1, \quad 0 \leqslant \beta_2 \leqslant \alpha_2, \quad \ldots, \quad 0 \leqslant \beta_r \leqslant \alpha_r.$$

Докажите это утверждение. В качестве следствия установите, что число всех делителей a (включая 1 и само a) равно произведению

$$(\alpha_1 + 1)(\alpha_2 + 1)\ldots(\alpha_r + 1).$$

Так, например,

$$144 = 2^4 \cdot 3^2$$

имеет $5 \cdot 3$ делителей. Вот они: 1, 2, 4, 8, 16, 3, 6, 12, 24, 48, 9, 18, 36, 72, 144.

2. Распределение простых чисел. Можно составить список всех простых чисел, не превышающих какого-то данного числа N, следующим образом. Напишем подряд все натуральные числа от 2 до N, затем вычеркнем все числа, являющиеся кратными 2 (не считая самого числа 2), все числа, являющиеся кратными 3 (не считая 3), и т. д., пока не будут вычеркнуты все составные числа. Эта процедура, известная под названием «решета Эратосфена», позволит выловить все простые числа в пределах от 2 до N. Усовершенствования этого метода мало-помалу привели к тому, что в настоящее время составлены вполне надежные таблицы простых чисел примерно до 10 000 000. Они предоставляют в наше распоряжение обширнейший эмпирический материал, позволяющий судить о распределении и свойствах простых чисел. Основываясь на этих таблицах, мы можем высказать ряд в высшей степени правдоподобных гипотез — совершенно так, как будто бы теория чисел была экспериментальной наукой. Часто доказательство этих гипотез оказывается необычайно затруднительным.

а. Формулы, дающие простые числа

Были сделаны попытки найти элементарные арифметические формулы, которые давали бы только простые числа, хотя бы без требования, чтобы они давали *все* простые числа. Ферма высказал предположение (не выставляя его в качестве положительного утверждения), что все числа вида

$$F(n) = 2^{2^n} + 1$$

являются простыми. В самом деле, при $n = 1, 2, 3, 4$ мы получаем

$$F(1) = 2^2 + 1 = 5,$$
$$F(2) = 2^{2^2} + 1 = 2^4 + 1 = 17,$$
$$F(3) = 2^{2^3} + 1 = 2^8 + 1 = 257,$$
$$F(4) = 2^{2^4} + 1 = 2^{16} + 1 = 65\,537$$

— всё простые числа. Но в 1732 г. Эйлер разложил на множители число $2^{2^5} + 1 = 641 \cdot 6\,700\,417$; таким образом, число $F(5)$ — уже не простое. Позднее среди этих «чисел Ферма» удалось обнаружить другие составные числа, причем ввиду непреодолимых трудностей, с которыми были связаны непосредственные пробы, в каждом случае были выработаны более глубокие теоретико-числовые методы. В настоящее время остается неизвестным даже то, дает ли формула Ферма бесконечное множество простых чисел.

Вот другое простое и замечательное выражение, дающее много простых чисел:

$$f(n) = n^2 - n + 41.$$

При $n = 1, 2, 3, \ldots, 40$ $f(n)$ есть простое число; но уже при $n = 41$ простого числа не получается:

$$f(41) = 41^2.$$

Выражение

$$n^2 - 79n + 1601$$

дает простые числа до $n = 79$ включительно; при $n = 80$ получается составное число.

В итоге можно сказать, что поиски элементарных формул, дающих только простые числа, оказались тщетными. Еще менее обнадеживающей следует считать задачу нахождения такой формулы, которая давала бы только простые числа и притом все простые числа.

б. Простые числа в арифметических прогрессиях

Если доказательство того, что в последовательности всех натуральных чисел $n = 1, 2, 3, 4, \ldots$ содержится бесконечное множество простых чисел, носит вполне элементарный характер, то следующий шаг в сторону таких последовательностей, как, например, 1, 4, 7, 10, 13, … или 3, 7, 11, 15, 19, …, или, вообще говоря, в сторону произвольной арифметической прогрессии a, $a + d$, $a + 2d$, …, $a + nd$, … (где a и d не имеют общих множителей), оказался связанным с гораздо бо́льшими трудностями. Все наблюдения только подтверждали тот факт, что *в каждой такой прогрессии содержится бесконечное число простых чисел*, как и в простейшей из них 1, 2, 3, … Но понадобились величайшие усилия для того, чтобы доказать эту общую теорему. Успех был достигнут Лежёном Дирихле (1805–1859), одним из ведущих математиков прошлого столетия, который применил при доказательстве самые усовершенствованные средства математического анализа из известных в то время. Его замечательные работы в этой области даже для настоящего времени остаются непревзойденными; прошло около ста лет, но доказательства Дирихле все еще не упрощены настолько, чтобы они могли быть поняты теми, кто не овладел полностью техникой математического анализа и теорией функций.

Мы не будем здесь пытаться привести доказательство общей теоремы Дирихле, а ограничимся рассмотрением более легкой задачи: обобщим евклидово доказательство о существовании бесконечного множества простых чисел таким образом, чтобы оно охватило некоторые *специальные* прогрессии, например $4n + 3$ или $6n + 5$. Рассмотрим первую из этих прогрессий. Заметим прежде всего, что всякое простое число, большее 2, — непременно нечетное (иначе оно делилось бы на 2)

и, следовательно, имеет вид $4n+1$ или $4n+3$ (при целом n). Далее, произведение двух чисел вида $4n+1$ также есть число того же вида, так как

$$(4a+1)(4b+1) = 16ab + 4a + 4b + 1 = 4(4ab + a + b) + 1.$$

Допустим теперь, что существует лишь конечное число простых чисел вида $4n+3$; обозначим их $p_1, p_2, \ldots, p_n$ и рассмотрим число

$$N = 4(p_1 p_2 \ldots p_n) - 1 = 4(p_1 \ldots p_n - 1) + 3.$$

Одно из двух: либо число N — простое, либо оно разлагается в произведение простых чисел, среди которых, однако, не может быть ни одного из чисел $p_1, p_2, \ldots, p_n$, так как эти числа делят N с остатком -1.

Заметим далее, что все множители, входящие в N, не могут быть вида $4n+1$, так как само N не этого вида, а мы видели, что произведение чисел вида $4n+1$ является числом того же вида. Итак, хоть один из множителей, входящих в N, должен быть вида $4n+3$, а это невозможно, так как ни одно из чисел p не входит множителем в N, а числами p все простые числа вида $4n+3$ по предположению исчерпываются. Таким образом, допуская, что существует лишь конечное число простых чисел вида $4n+3$, мы приходим к противоречию, и значит, таких чисел бесконечно много.

Упражнение. Докажите соответствующую теорему для прогрессии $6n + 5$.

в. Теорема о распределении простых чисел

В исследованиях, связанных с законом распределения простых чисел, решительный шаг был сделан тогда, когда математики отказались от тщетных попыток найти элементарную математическую формулу, которая давала бы все простые числа или же точное число простых чисел, содержащихся среди n первых натуральных чисел, и сосредоточили вместо того внимание на распределении в *среднем* простых чисел среди всех натуральных.

При всяком целом n обозначим через A_n число простых чисел среди чисел $1, 2, 3, \ldots, n$. Если мы выделим среди первых чисел натурального ряда

1 **2** **3** 4 **5** 6 **7** 8 9 10 **11** 12 **13** 14 15 16 **17** 18 **19** ...

те, которые являются простыми, то не составит труда подсчитать ряд значений A_n:

$$A_1 = 0, \quad A_2 = 1, \quad A_3 = A_4 = 2, \quad A_5 = A_6 = 3,$$
$$A_7 = A_8 = A_9 = A_{10} = 4, \quad A_{11} = A_{12} = 5,$$
$$A_{13} = A_{14} = A_{15} = A_{16} = 6, \quad A_{17} = A_{18} = 7, \quad A_{19} = 8 \quad \text{и т. д.}$$

Возьмем теперь какую-нибудь неограниченно возрастающую последовательность значений n, например

$$n = 10,\ \ 10^2,\ \ 10^3,\ \ 10^4,\ \ \ldots;$$

тогда соответствующие значения A_n

$$A_{10},\ \ A_{10^2},\ \ A_{10^3},\ \ A_{10^4},\ \ \ldots$$

также будут возрастать безгранично (хотя и более медленно). Действительно, множество простых чисел, как мы уже знаем, бесконечно, и потому значения A_n рано или поздно станут больше любого назначенного числа. «Плотность» распределения простых чисел среди n первых чисел натурального ряда дается отношением $\frac{A_n}{n}$; не представляет особого труда с помощью таблиц простых чисел подсчитать значения $\frac{A_n}{n}$ при достаточно больших значениях n:

n	$\frac{A_n}{n}$
10^3	0,168
10^6	0,078498
10^9	0,050847478

Последняя, скажем, из выписанных строчек в приведенной табличке дает *вероятность* того, что число, случайно выхваченное из 10^9 первых чисел натурального ряда, окажется простым: всего имеется 10^9 возможных выборов, из них A_{10^9} соответствуют простым числам.

Распределение отдельных простых чисел отличается чрезвычайно неправильным характером. Но эта неправильность «в малом» исчезает, если мы направим внимание к распределению «в среднем», находящему свое выражение в изменениях отношения $\frac{A_n}{n}$ при неограниченно растущем n. Простой закон, которому подчиняется поведение этого отношения, следует отнести к числу самых замечательных открытий, сделанных во всей математике. Для того чтобы сформулировать *теорему о распределении простых чисел*, к которой мы теперь подходим, необходимо предварительно разъяснить, что такое «натуральный логарифм» числа n. Для этой цели возьмем в плоскости две взаимно перпендикулярные оси и рассмотрим геометрическое место таких точек на плоскости, для которых произведение расстояний x и y от двух осей

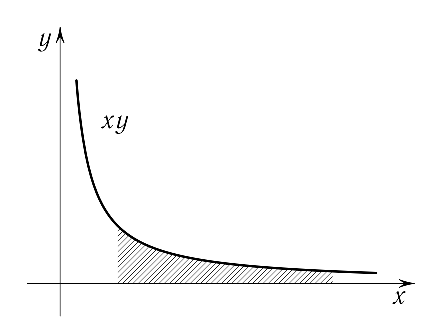


Рис. 5. Площадь заштрихованной

равно единице. В терминах координат x и y это геометрическое место есть равносторонняя гипербола, уравнение которой имеет вид $xy = 1$. Мы определим $\ln n$ как *площадь* (рис. 5) фигуры, ограниченной гиперболой и двумя вертикальными прямыми $x = 1$ и $x = n$. (Более детально логарифм и его свойства будут рассмотрены в главе VIII.) Чисто случайно, в связи с изучением таблицы простых чисел, Гаусс заметил, что отношение $\frac{A_n}{n}$ приблизительно равно $\frac{1}{\ln n}$ и что точность этого приближения, по-видимому, улучшается при возрастании n. Насколько удовлетворительно приближение, можно судить по отношению $\frac{A_n}{n} : \frac{1}{\ln n}$, значения которого при $n = 1000$, $1\,000\,000$, $1\,000\,000\,000$ показаны в следующей табличке:

n	$\frac{A_n}{n}$	$\frac{1}{\ln n}$	$\frac{A_n}{n} : \frac{1}{\ln n}$
10^3	0,168	0,145	1,159
10^6	0,078498	0,072382	1,084
10^9	0,050847478	0,048254942	1,053
....			....

Основываясь на такого рода эмпирической очевидности, Гаусс высказал в качестве предположения, что отношение $\frac{A_n}{n}$ «асимптотически равно» $\frac{1}{\ln n}$. Смысл этого утверждения заключается в следующем: если возьмем последовательность значений n, становящихся все больше и больше, например,

$$10,\ 10^2,\ 10^3,\ 10^4,\ \ldots$$

(как мы делали и раньше), то отношение

$$\frac{A_n}{n} : \frac{1}{\ln n},$$

вычисляемое для этих последовательно рассматриваемых значений n, будет становиться все более и более близким к числу 1, а именно, разность между указанным отношением и единицей будет делаться столь малой, сколь будет назначено, лишь бы только мы рассматривали достаточно большие значения n. Такого рода соотношение символически выражается знаком $\sim$: $\frac{A_n}{n} \sim \frac{1}{\ln n}$ означает, что $\frac{A_n}{n} : \frac{1}{\ln n}$ при возрастании n стремится к 1. Что знак $\sim$ не может быть заменен знаком обыкновенного равенства ($=$), ясно хотя бы из того факта, что A_n — непременно целое число, тогда как $\frac{n}{\ln n}$ не является таковым.

То обстоятельство, что распределение простых чисел хорошо описывается с помощью логарифмической функции, нельзя не признать поистине поразительным, так как здесь вступают в тесное соприкосновение два математических понятия, казалось бы не имеющие друг к другу никакого отношения.

Хотя схватить содержание высказанного Гауссом предположения не представляет особой трудности, однако его строгое математическое доказательство во времена Гаусса было за пределами возможностей математической науки. Для того чтобы доказать теорему о распределении простых чисел, говорящую лишь о самых элементарных математических понятиях, неизбежно нужно прибегнуть к самым мощным методам современной математики. Пришлось ждать почти сто лет, пока анализ получил достаточное развитие для того, чтобы Адамар (1896) в Париже и Валле-Пуссен (1896) в Лувэне смогли дать исчерпывающее доказательство теоремы о распределении простых чисел. Упрощения и важные дополнения были затем внесены Мангольдтом и Э. Ландау. Задолго до Адамара значительное продвижение в этой области было сделано Риманом (1826–1866) в его знаменитой работе, намечающей основные стратегические линии предстоящей атаки. Совсем недавно американский математик Норберт Винер сумел видоизменить доказательство таким образом, чтобы избежать применения комплексных чисел в узловых моментах проводимых рассуждений. Но все же доказательство теоремы о распределении простых чисел остается слишком сложным для того, чтобы его можно было предложить начинающему. Мы вернемся к этому вопросу на стр. 507 и следующих.

г. Две еще не решенные задачи о простых числах

Если проблема распределения простых чисел («в среднем») была разрешена удовлетворительно, то справедливость ряда других гипотез, эмпирически совершенно несомненная, все еще не доказана.

Сюда относится прежде всего знаменитая *гипотеза Гольдбаха.* Гольдбах (1690–1764) сам по себе не оставил никакого следа в истории математики: он прославился только проблемой, которую предложил Эйлеру в письме, относящемся к 1742 г. Он обратил внимание на тот факт, что ему всегда удавалось представить любое четное число (кроме 2, которое само есть простое число) в виде суммы двух простых. Например, $4 = 2 + 2$, $6 = 3 + 3$, $8 = 5 + 3$, $10 = 5 + 5$, $12 = 5 + 7$, $14 = 7 + 7$, $16 = 13 + 3$, $18 = 11 + 7$, $20 = 13 + 7$, ..., $48 = 29 + 19$, ..., $100 = 97 + 3$ и т. д.

Гольдбах спрашивал у Эйлера, может ли тот доказать, что такого рода представление возможно для всякого четного числа, или же, напротив, сможет указать пример, опровергающий такое предположение. Эйлер так и не дал ответа; не дал его никто и в дальнейшем.

Эмпирическая очевидность гипотезы Гольдбаха, как легко проверить, вполне убедительна. Источник же возникающих затруднений — в том, что понятие простого числа определяется в терминах *умножения*, тогда как сама проблема касается *сложения*. Вообще, находить связи между мультипликативными и аддитивными свойствами чисел очень трудно.

До недавнего времени доказательство гипотезы Гольдбаха казалось задачей совершенно неприступной. Сегодня дело обстоит уже не так. Очень значительный успех, оказавшийся неожиданным и поразительным для всех специалистов по данному вопросу, был достигнут в 1931 г. неизвестным в то время молодым русским математиком Шнирельманом (1905–1938), который доказал, что *всякое целое положительное число может быть представлено в виде суммы не более чем* 800 000 *простых*. Хотя этот результат и производит несколько комическое впечатление (по сравнению с первоначально поставленной целью доказать гипотезу Гольдбаха), тем не менее он стал первым шагом в должном направлении. Доказательство Шнирельмана — прямое и носит конструктивный характер, хотя и не обеспечивает практического метода для представления произвольного целого числа в виде суммы простых. Еще позднее русский же математик Виноградов, пользуясь методами Харди, Литтлвуда и их поистине великого сотрудника по работе индуса Рамануджана, сумел понизить число слагаемых в формулировке Шнирельмана с 800 000 до 4. Это уже гораздо ближе к решению проблемы Гольдбаха. Но между результатами Шнирельмана и Виноградова имеется очень резкое различие более резкое, чем различие между числами 800 000 и 4. Теорема Виноградова была доказана им лишь для всех «достаточно больших» чисел; точнее говоря, Виноградов установил *существование* такого числа N, что всякое целое число $n > N$ может быть представлено в виде суммы четырех простых чисел. Метод Виноградова не позволяет никак судить о величине N; в противоположность методу Шнирельмана, он — существенно «косвенный» и неконструктивный. По существу, Виноградов доказал следующее: допуская, что существует *бесконечное множество чисел, не представимых в виде суммы четырех* (*или менее того*) *простых чисел*, можно получить противоречие. Здесь перед нами прекрасный пример, показывающий глубокое различие между двумя типами доказательств — прямым и косвенным (см. общее обсуждение этого вопроса на стр. 40)[1].

[1] Основной результат И. М. Виноградова (1937) устанавливает существование такого натурального N, что всякое *нечетное* $n > N$ представимо в виде суммы трех простых чисел:

$$n = p_1 + p_2 + p_3, \qquad (*)$$

из чего, разумеется, вытекает уже представимость *любого* натурального $n > N + 2$ в виде суммы *четырех* простых чисел. Результат Виноградова о нечетных числах окончателен — число слагаемых (три) в его формулировке уменьшить нельзя. Что же касается *четных* чисел, то из представимости их в виде (*)

Следующая проблема, еще более любопытная, чем проблема Гольдбаха, нисколько не приблизилась к своему разрешению. Было подмечено, что простые числа нередко встречаются парами в виде p и $p+2$. Таковы 3 и 5, 11 и 13, 29 и 31 и т. д. Предположение о существовании бесконечного множества таких «близнецов» кажется весьма правдоподобным, но до сих пор не удалось даже приблизиться к его доказательству.

§ 2. Сравнения

1. Общие понятия. Всякий раз, когда приходится говорить о делимости целых чисел на некоторое определенное целое число d, все рассуждения становятся яснее и проще, если пользоваться отношением сравнения, введенным Гауссом, и соответствующими обозначениями.

Чтобы ввести понятие сравнения, рассмотрим остатки, получающиеся при делении различных чисел, например, на 5. Мы получаем:

$$\begin{array}{lll}
0 = 0\cdot 5+0 & 7 = 1\cdot 5+2 & -1 = -1\cdot 5+4\\
1 = 0\cdot 5+1 & 8 = 1\cdot 5+3 & -2 = -1\cdot 5+3\\
2 = 0\cdot 5+2 & 9 = 1\cdot 5+4 & -3 = -1\cdot 5+2\\
3 = 0\cdot 5+3 & 10 = 2\cdot 5+0 & -4 = -1\cdot 5+1\\
4 = 0\cdot 5+4 & 11 = 2\cdot 5+1 & -5 = -1\cdot 5+0\\
5 = 1\cdot 5+0 & 12 = 2\cdot 5+2 & -6 = -2\cdot 5+4\\
6 = 1\cdot 5+1 & \text{и т. д.} & \text{и т. д.}
\end{array}$$

Заметим, что остатком при делении на 5 может быть только одно из чисел 0, 1, 2, 3, 4. Говорят, что два числа a и b *сравнимы по модулю* 5, если при делении на 5 они дают *один и тот же остаток*. Так, все числа 2, 7, 12, 17, 22, ..., −3, −8, −13, −18, ... сравнимы по модулю 5, так как при делении на 5 все они дают остаток 2. Вообще, говорят, что два числа a и b *сравнимы по модулю* d (где d — некоторое целое число), если при делении на d они дают один и тот же остаток; другими словами, если существует такое целое число n (положительное, отрицательное или

вытекала бы сразу и представимость любого четного n в виде суммы двух простых слагаемых (т. к. в этом случае одно из слагаемых равно 2). Однако при всей правдоподобности гипотезы о представимости в таком виде любого четного $n > 3$, проблема ее доказательства чрезвычайно трудна и превышает, по-видимому, еще возможности математиков.

Неэффективный характер теоремы Виноградова устранен в 1939 г. К. Г. Бороздиным, показавшим, что в виде (*) представимо любое нечетное $n > C = e^{e^{e^{41,96}}}$; в 1956 г. ему же удалось снизить эту оценку до $C = e^{e^{17}}$. Конечно, уменьшение константы C до разумных пределов позволило бы решить гипотезу о представимости в виде (*) нечетных n, $6 < n < C$, — и тем самым любого нечетного $n > 6$ — посредством прямой вычислительной проверки. — *Прим. А. Н. Колмогорова.*

нуль), что $a - b = nd$. Например, 27 и 15 сравнимы по модулю 4, так как $27 = 6 \cdot 4 + 3$, $15 = 3 \cdot 4 + 3$.

Для отношения сравнения введено специальное обозначение — если a и b сравнимы по модулю d, то пишут: $a \equiv b \pmod{d}$. [Если же a не сравнимо с b по модулю d, то пишут $a \not\equiv b \pmod{d}$.] Если ясно, какой модуль имеется в виду, то приписку «mod d» опускают.

Сравнения часто встречаются в повседневной жизни. Например, часовая стрелка указывает время по модулю 12; автомобильный счетчик отмечает пройденные расстояния по модулю 100 000 (миль или километров).

Прежде чем перейти к более детальному рассмотрению сравнений и их свойств, пусть читатель проверит, что следующие утверждения в точности эквивалентны:

1) a сравнимо с b по модулю d.
2) $a = b + nd$, где n — целое.
3) $a - b$ делится на d.

Введенные Гауссом обозначения для сравнений подчеркивают то обстоятельство, что сравнения обладают многими свойствами обычных равенств. Напомним эти свойства:

1) Всегда $a = a$.
2) Если $a = b$, то $b = a$.
3) Если $a = b$ и $b = c$, то $a = c$.

Кроме того, если $a = a'$ и $b = b'$, то

4) $a + b = a' + b'$.
5) $a - b = a' - b'$.
6) $ab = a'b'$.

Эти же свойства сохраняются, если соотношение равенства $a = b$ заменяется соотношением сравнения $a \equiv b \pmod{d}$. Именно:

1′) Всегда $a \equiv a \pmod{d}$.
2′) Если $a \equiv b \pmod{d}$, то $b \equiv a \pmod{d}$.
3′) Если $a \equiv b \pmod{d}$ и $b \equiv c \pmod{d}$, то $a \equiv c \pmod{d}$.

(Проверьте! — это нетрудно.)
Точно так же, если $a \equiv a' \pmod{d}$ и $b \equiv b' \pmod{d}$, то

4′) $a + b \equiv a' + b' \pmod{d}$.
5′) $a - b \equiv a' - b' \pmod{d}$.
6′) $ab \equiv a'b' \pmod{d}$.

Таким образом, *сравнения по одному и тому же модулю можно складывать, вычитать и умножать*. В самом деле, из

$$a = a' + rd, \qquad b = b' + sd$$

вытекает

$$a + b = a' + b' + (r + s)d,$$
$$a - b = a' - b' + (r - s)d,$$
$$ab = a'b' + (a's + b'r + rsd)d,$$

что и приводит к нужным заключениям.

Сравнения допускают великолепное геометрическое представление. Если хотят дать геометрическое представление целым числам, то обыкновенно выбирают прямолинейный отрезок единичной длины и затем откладывают кратные отрезки в обе стороны. Таким образом, для каждого целого числа получается соответствующая ему точка на прямой — числовой оси (рис. 6). Но если приходится иметь

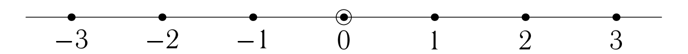


Рис. 6. Геометрическое представление целых чисел

дело с числами по данному модулю d, два сравнимых числа — поскольку речь идет о делимости на d — рассматриваются как нечто неразличимое, так как дают одни и те же остатки. Чтобы изобразить все это геометрически, возьмем окружность, разделенную на d равных частей. Всякое целое число при делении на d дает в качестве остатка одно из чисел $0, 1, 2, \ldots, d - 1$; эти числа мы и расставим по окружности на равных расстояниях. Каждое число сравнимо с одним из этих чисел по модулю d и, следовательно, представляется соответствующей точкой; два числа сравнимы, если изображаются одной и той же точкой. Рис. 7 сделан для случая $d = 6$. Циферблат часов может также служить моделью.

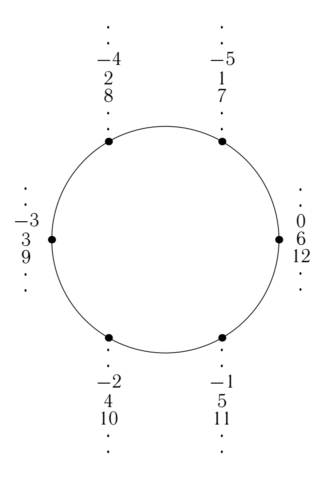


Рис. 7. Геометрическое представление целых чисел по модулю 6

В качестве примера применения мультипликативного свойства сравнений 6′) определим остатки, получающиеся при делении на одно и то же число последовательных степеней числа 10. Так как $10 = -1 + 11$, то

$$10 \equiv -1 \pmod{11}.$$

Умножая многократно это сравнение само на себя, получаем дальше

$$10^2 \equiv (-1)(-1) = 1 \pmod{11},$$
$$10^3 \equiv (-1) \pmod{11},$$
$$10^4 \equiv 1 \pmod{11} \quad \text{и т. д.}$$

Отсюда можно заключить, что всякое целое число, запись по десятичной

системе которого имеет вид

$$z = a_0 + a_1 \cdot 10 + a_2 \cdot 10^2 + \ldots + a_n \cdot 10^n,$$

дает тот же остаток при делении на 11, что и сумма его цифр, взятая с чередующимися знаками:

$$t = a_0 - a_1 + a_2 - \ldots$$

В самом деле, мы имеем

$$z - t = a_1 \cdot 11 + a_2(10^2 - 1) + a_3(10^3 + 1) + a_4(10^4 - 1) + \ldots$$

Так как все выражения $10^2 - 1$, $10^3 + 1$, $\ldots$ сравнимы с нулем по модулю 11, то $z - t$ также сравнимо с нулем, и потому z при делении на 11 дает тот же остаток, что и t. В частности, число делится на 11, т. е. дает остаток 0 при делении, в том и только том случае, если знакочередующаяся сумма его цифр делится на 11. Например, число $z = 3\,162\,819$ делится на 11, так как $3 - 1 + 6 - 2 + 8 - 1 + 9 = 22$ делится на 11. Найти таким же образом правило делимости на 3 или на 9 еще проще, так как $10 \equiv 1$ (mod 3 и 9), и потому $10^n \equiv 1$ (mod 3 и 9) при любом n. Отсюда следует, что число z делится на 3 и на 9 в том и только том случае, если сумма его цифр

$$s = a_0 + a_1 + a_2 + \ldots + a_n$$

делится соответственно на 3 и на 9.

Если в качестве модуля возьмем 7, то получим

$$10 \equiv 3, \quad 10^2 \equiv 2, \quad 10^3 \equiv -1, \quad 10^4 \equiv -3, \quad 10^5 \equiv -2, \quad 10^6 \equiv 1.$$

Далее остатки повторяются. Таким образом, z делится на 7 в том и только том случае, если выражение

$$r = a_0 + 3a_1 + 2a_2 - a_3 - 3a_4 - 2a_5 + a_6 + 3a_7 + \ldots$$

делится на 7.

Упражнение. Найдите подобный же признак делимости на 13.

Складывая и умножая сравнения по определенному модулю, скажем $d = 5$, можно всегда обеспечить то, чтобы входящие числа не становились слишком большими, заменяя всякий раз встречающееся число одним из чисел

$$0,\ 1,\ 2,\ 3,\ 4,$$

а именно тем, с которым оно сравнимо. Так, вычисляя суммы и произведения различных чисел по модулю 5, нужно только пользоваться следующими таблицами сложения и умножения:

$a+b$

	$b\equiv 0$	1	2	3	4
$a\equiv 0$	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

ab

	$b\equiv 0$	1	2	3	4
$a\equiv 0$	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Из второй таблицы видно, что произведение ab сравнимо с нулем по модулю 5 только в том случае, если a или $b\equiv 0 \pmod 5$. Это наводит на мысль о существовании следующего общего закона:

7) $ab\equiv 0 \pmod d$ только в том случае, если $a\equiv 0$ или $b\equiv 0 \pmod d$,

что является распространением хорошо известного свойства обыкновенного умножения:

$ab=0$ только в том случае, если $a=0$ или $b=0$.

Но закон 7) *действителен только при том условии, что модуль d есть простое число.* Действительно, сравнение

$$ab\equiv 0 \pmod d$$

означает, что ab делится на d, а мы уже видели, что произведение ab делится на простое d в том и только том случае, если один из множителей a или b делится на d, т. е. если

$$a\equiv 0 \pmod d \quad \text{или} \quad b\equiv 0 \pmod d.$$

С другой стороны, закон теряет силу при d составном: можно тогда написать $d=r\cdot s$, где оба множителя r и s меньше чем d, так что

$$r\not\equiv 0 \pmod d, \qquad s\not\equiv 0 \pmod d$$

и, однако,

$$rs=d\equiv 0 \pmod d.$$

Например, $2\not\equiv 0 \pmod 6$ и $3\not\equiv 0 \pmod 6$, но $2\cdot 3=6\equiv 0 \pmod 6$.

Упражнения. 1) Покажите, что для сравнений по простому модулю имеет место следующее *правило сокращения*: если $ab\equiv ac$ и $a\not\equiv 0$, то $b\equiv c$.

2) С каким числом в пределах от 0 до 6 включительно сравнимо число $11\cdot 18\times 2322\cdot 13\cdot 19$ по модулю 7?

3) С каким числом в пределах от 0 до 12 включительно сравнимо число $3\cdot 7\times 11\cdot 17\cdot 19\cdot 23\cdot 29\cdot 113$ по модулю 13?

4) С каким числом в пределах от 0 до 4 включительно сравнима сумма $1+2+2^2+\ldots+2^{19}$ по модулю 5?

2. Теорема Ферма. В XVII столетии Ферма, основатель современной теории чисел, открыл чрезвычайно важную теорему. *Если p — простое число, не делящее целого числа a, то*

$$a^{p-1}\equiv 1 \pmod p.$$

Другими словами, $(p-1)$-я степень a при делении на p дает остаток 1.

Некоторые из ранее произведенных нами вычислений подтверждают эту теорему: так, мы видим, что $10^6 \equiv 1 \pmod 7$, $10^2 \equiv 1 \pmod 3$ и $10^{10} \equiv 1 \pmod{11}$. Таким же образом легко проверить, что $2^{12} \equiv 1 \pmod{13}$ и $5^{10} \equiv 1 \pmod{11}$. Для этой цели нет необходимости на самом деле вычислять столь высокие степени данных чисел; достаточно использовать мультипликативное свойство сравнений:

$$\begin{aligned} &2^4 = 16 \equiv 3 \pmod{13}, && 5^2 \equiv 3 \pmod{11},\\ &2^8 \equiv 9 \equiv -4 \pmod{13}, && 5^4 \equiv 9 \equiv -2 \pmod{11},\\ &2^{12} \equiv -4\cdot 3 = -12 \equiv 1 \pmod{13}, && 5^8 \equiv 4 \pmod{11},\\ & && 5^{10} \equiv 3\cdot 4 = 12 = 1 \pmod{11}. \end{aligned}$$

Обращаясь к доказательству теоремы Ферма, рассмотрим числа, кратные a:

$$m_1 = a, \quad m_2 = 2a, \quad m_3 = 3a, \ \ldots, \ m_{p-1} = (p-1)a.$$

Никакие два из этих чисел не могут быть между собой сравнимы по модулю p. В противном случае p должно было бы делить разность $m_r - m_s = (r-s)a$, где r, s была бы пара целых чисел, подчиненных ограничению $1 \leqslant r < s \leqslant (p-1)$. Но из закона 7) следует, что этого не может случиться: так как $s-r$ меньше чем p, то p не делит $s-r$; с другой стороны, по предположению, p не делит и a. Таким же образом мы убеждаемся, что ни одно из чисел m не сравнимо с нулем. Отсюда следует, что числа $m_1, m_2, \ldots, m_{p-1}$ соответственно сравнимы с числами 1, 2, $\ldots$, $p-1$, взятыми в некоторой их перестановке. Дальше заключаем:

$$m_1 m_2 \ldots m_{p-1} = 1\cdot 2\cdot 3 \ldots (p-1)a^{p-1} \equiv 1\cdot 2\cdot 3 \ldots (p-1) \pmod p,$$

или же, полагая ради краткости $K = 1\cdot 2\cdot 3 \ldots (p-1)$,

$$K(a^{p-1}-1) \equiv 0 \pmod p.$$

Число K не делится на p, так как ни один из входящих в него множителей не делится на p; значит, согласно закону 7) $(a^{p-1}-1)$ должно делиться на p, т. е.

$$a^{p-1} - 1 \equiv 0 \pmod p.$$

Это и есть теорема Ферма.

Проверим эту теорему еще раз. Возьмем $p = 23$ и $a = 5$; тогда получаем по модулю 23

$$5^2 \equiv 2, \quad 5^4 \equiv 4, \quad 5^8 \equiv 16 = -7, \quad 5^{16} \equiv 49 \equiv 3, \quad 5^{20} \equiv 12, \quad 5^{22} \equiv 24 \equiv 1.$$

Если возьмем $a = 4$ вместо 5, то будем иметь, опять-таки по модулю 23,

$$4^2 \equiv -7, \ \ 4^3 \equiv -28 \equiv -5, \ \ 4^4 \equiv -20 = 3, \ \ 4^8 \equiv 9, \ \ 4^{11} \equiv -45 \equiv 1, \ \ 4^{22} \equiv 1.$$

В примере, где было взято $a = 4$, $p = 23$ (как и во многих иных), можно заметить, что не только $(p-1)$-я степень, но и более низкая степень a

уже оказывается сравнимой с единицей. Наименьшая такая степень — в нашем примере степень 11 — *непременно* есть делитель числа $p-1$ (см. ниже, упражнение 3).

Упражнения. 1) С помощью подобных же вычислений проверьте, что

$$2^8 \equiv 1 \pmod{17}, \qquad 3^8 \equiv -1 \pmod{17}, \qquad 3^{14} \equiv -1 \pmod{29},$$
$$2^{14} \equiv -1 \pmod{29}, \qquad 4^{14} \equiv 1 \pmod{29}, \qquad 5^{14} \equiv 1 \pmod{29}.$$

2) Проверьте теорему Ферма, взяв $p = 5$, 7, 11, 17 и 23 и придавая числу a различные значения.

3) Докажите общую теорему: *наименьшее число e, для которого $a^e \equiv 1 \pmod{p}$, должно быть делителем $p-1$.* [*Указание*: произведите деление $p-1$ на e, получая

$$p - 1 = ke + r,$$

где $0 \leqslant r < e$, и дальше воспользуйтесь тем обстоятельством, что $a^{p-1} \equiv a^e \equiv 1 \pmod{p}$.]

3. Квадратические вычеты. Обращаясь снова к примерам, иллюстрирующим теорему Ферма, мы можем подметить, что не только всегда справедливо сравнение $a^{p-1} \equiv 1 \pmod{p}$, но (предположим, что p есть простое число, отличное от 2, значит, — нечетное, $p = 2p' + 1$) при некоторых значениях a справедливо также сравнение $a^{p'} = a^{\frac{p-1}{2}} \equiv 1 \pmod{p}$. Это обстоятельство вызывает ряд заслуживающих внимания соображений. Теорему Ферма можно записать в следующем виде:

$$a^{p-1} - 1 = a^{2p'} - 1 = (a^{p'} - 1)(a^{p'} + 1) \equiv 0 \pmod{p}.$$

Так как произведение делится на p только в том случае, если один из множителей делится на p, то, значит, одно из чисел $a^{p'} - 1$ или $a^{p'} + 1$ должно делиться на p; поэтому, каково бы ни было простое число $p > 2$ и каково бы ни было число a, не делящееся на p, непременно должно иметь место одно из двух сравнений

$$a^{\frac{p-1}{2}} \equiv 1 \qquad \text{или} \qquad a^{\frac{p-1}{2}} \equiv -1 \pmod{p}.$$

Начиная с самого возникновения современной теории чисел, математики были заинтересованы выяснением вопроса: для каких чисел a оправдывается первое сравнение, а для каких — второе? Предположим, что a сравнимо по модулю p с квадратом некоторого числа x,

$$a \equiv x^2 \pmod{p}.$$

Тогда $a^{\frac{p-1}{2}} \equiv x^{p-1}$, и согласно теореме Ферма правая, а следовательно, и левая части сравнения должны быть сравнимы с 1 по модулю p. Такое число a (не являющееся кратным p), которое по модулю p сравнимо с квадратом некоторого числа, называется *квадратическим вычетом p*; напротив, число b, не кратное p, которое не сравнимо ни с каким квадратом по модулю p, называется *квадратическим невычетом p*. Мы только

что видели, что всякий квадратический вычет a числа p удовлетворяет сравнению $a^{\frac{p-1}{2}} \equiv 1 \pmod p$. Довольно легко установить, что всякий невычет b числа p удовлетворяет сравнению $b^{\frac{p-1}{2}} \equiv -1 \pmod p$. Кроме того, мы покажем (несколько дальше), что среди чисел $1, 2, 3, \ldots, p-1$ имеется в точности $\frac{p-1}{2}$ квадратических вычетов и $\frac{p-1}{2}$ невычетов.

Хотя с помощью прямых подсчетов можно было собрать немало эмпирических данных, но открыть сразу общие законы, регулирующие распределение квадратических вычетов, было нелегко. Первое глубоко лежащее свойство этих вычетов было подмечено Лежандром (1752–1833); позднее Гаусс назвал его *законом взаимности*. Этот закон касается взаимоотношения между двумя различными простыми числами p и q. Он заключается в следующем:

1) Предположим, что произведение $\frac{p-1}{2} \cdot \frac{q-1}{2}$ четное. Тогда q есть вычет p в том и только том случае, если p есть вычет q.

2) Предположим, напротив, что указанное произведение — *нечетное*. Тогда ситуация резко меняется: q есть вычет p, если p есть *невычет* q, и наоборот.

Первое строгое доказательство закона взаимности, долгое время остававшегося гипотезой, данное Гауссом еще в молодости, явилось одним из крупных его достижений. Доказательство Гаусса никоим образом нельзя назвать простым, и в наше время провести доказательство закона взаимности стоит известного труда, хотя количество различных опубликованных доказательств очень велико. Истинный смысл закона взаимности вскрылся лишь в недавнее время — в связи с новейшим развитием алгебраической теории чисел.

В качестве примера, иллюстрирующего распределение квадратических вычетов, возьмем $p = 7$. Так как по модулю 7

$$0^2 \equiv 0, \quad 1^2 \equiv 1, \quad 2^2 \equiv 4, \quad 3^2 \equiv 2, \quad 4^2 \equiv 2, \quad 5^2 \equiv 4, \quad 6^2 \equiv 1$$

и так как дальнейшие квадраты повторяют эту последовательность, то квадратическими вычетами числа 7 являются числа, сравнимые с 1, 2 и 4, а невычетами — числа, сравнимые с 3, 5 и 6. В общем случае квадратические вычеты p составляются из чисел, сравнимых с числами 1^2, $2^2, \ldots, (p-1)^2$. Но эти последние попарно сравнимы, так как

$$x^2 \equiv (p-x)^2 \pmod p \qquad (\text{например, } 2^2 \equiv 5^2 \pmod 7).$$

Действительно, $(p-x)^2 = p^2 - 2px + x^2 \equiv x^2 \pmod p$. Значит, половина чисел $1, 2, \ldots, p-1$ представляет собою квадратические вычеты числа p, а другая половина — невычеты.

Чтобы дать иллюстрацию также и закону взаимности, положим $p = 5$, $q = 11$. Так как $11 \equiv 1^2 \pmod 5$, то 11 есть квадратический вычет по модулю 5, и так как, кроме того, произведение $\frac{5-1}{2} \cdot \frac{11-1}{2}$

четное, то согласно закону взаимности 5 должно быть также квадратическим вычетом по модулю 11; и в самом деле, мы видим, что $5 \equiv 4^2 \pmod{11}$. С другой стороны, положим $p = 7$, $q = 11$. Тогда произведение $\frac{7-1}{2} \cdot \frac{11-1}{2}$ нечетно, и в этом случае 11 есть вычет по модулю 7 (так как $11 \equiv 2^2 \pmod 7$), а 7 — невычет по модулю 11.

Упражнения. 1) $6^2 = 36 \equiv 13 \pmod{23}$. Является ли 23 квадратическим вычетом по модулю 13?

2) Мы видели, что $x^2 \equiv (p-x)^2 \pmod p$. Покажите, что иного вида сравнений между числами $1^2, 2^2, 3^2, \ldots, (p-1)^2$ быть не может.

§ 3. Пифагоровы числа и большая теорема Ферма

Интересный вопрос из области теории чисел связан с теоремой Пифагора. Теорема эта, как известно, алгебраически выражается равенством

$$a^2 + b^2 = c^2, \tag{1}$$

где a и b — длины катетов, а c — длина гипотенузы. Проблема разыскания *всех* прямоугольных треугольников, стороны которых выражаются целыми числами, таким образом, эквивалентна проблеме нахождения всех решений (a, b, c) в целых числах уравнения (1). Каждая тройка целых чисел (a, b, c), удовлетворяющих этому уравнению, носит название *пифагоровой тройки.*

Все пифагоровы тройки могут быть найдены довольно просто. Пусть целые числа a, b и c образуют пифагорову тройку, т. е. связаны соотношением $a^2 + b^2 = c^2$. Положим ради краткости $\frac{a}{c} = x$, $\frac{b}{c} = y$. Тогда x и y — рациональные числа, связанные равенством $x^2 + y^2 = 1$. Из последнего следует: $y^2 = (1-x)(1+x)$ или $\frac{y}{1+x} = \frac{1-x}{y}$. Общее значение двух отношений в полученной пропорции есть число t, которое может быть представлено как отношение двух целых чисел $\frac{u}{v}$. Можно, далее, написать: $y = t(1+x)$ и $(1-x) = ty$, или же

$$tx - y = -t, \qquad x + ty = 1.$$

Из полученной системы уравнений немедленно следует, что

$$x = \frac{1-t^2}{1+t^2}, \qquad y = \frac{2t}{1+t^2}.$$

Подставляя $\frac{a}{c}$ и $\frac{b}{c}$ вместо x и y и $\frac{u}{v}$ вместо t, будем иметь

$$\frac{a}{c} = \frac{v^2-u^2}{u^2+v^2}, \qquad \frac{b}{c} = \frac{2uv}{u^2+v^2}.$$

Отсюда вытекает

$$\left.\begin{aligned} a &= (v^2-u^2)r, \\ b &= (2uv)r, \\ c &= (u^2+v^2)r, \end{aligned}\right\} \tag{2}$$

где r — некоторый рациональный множитель пропорциональности. Итак, если числа $(a,\ b,\ c)$ образуют пифагорову тройку, то они соответственно пропорциональны числам вида v^2-u^2, $2uv$, u^2+v^2. Обратно, легко проверить, что всякие три числа $(a,\ b,\ c)$, определенные равенствами вида (2), образуют пифагорову тройку, так как из равенств (2) следует

$$\begin{aligned} a^2 &= (u^4 - 2u^2v^2 + v^4)r^2, \\ b^2 &= (4u^2v^2)r^2, \\ c^2 &= (u^4 + 2u^2v^2 + v^4)r^2, \end{aligned}$$

так что $a^2+b^2=c^2$.

Этот результат можно несколько упростить. Из некоторой пифагоровой тройки $(a,\ b,\ c)$ легко выводится бесконечное множество других пифагоровых троек $(sa,\ sb,\ sc)$, каково бы ни было целое положительное s. Так, из (3, 4, 5) получаются (6, 8, 10), (9, 12, 15) и т. д. Такие тройки не являются существенно различными, так как соответствуют подобным треугольникам. Мы условимся говорить о *примитивной* пифагоровой тройке, если числа a, b и c не имеют общего множителя. Можно показать, что *формулы*

$$\begin{aligned} a &= v^2 - u^2, \\ b &= 2uv, \\ c &= u^2 + v^2, \end{aligned}$$

где u, v — произвольные целые положительные числа $(v > u)$, не имеющие общих множителей и не являющиеся одновременно нечетными, дают нам все примитивные пифагоровы тройки.

* **Упражнение.** Докажите последнее утверждение.

Вот примеры примитивных пифагоровых троек:

$v=2,\ u=1\quad (3,\ 4,\ 5)$, $\qquad v=4,\quad u=3\quad (7,\ 24,\ 25)$,

$v=3,\ u=2\quad (5,\ 12,\ 13)$, $\qquad v=10,\quad u=7\quad (51,\ 140,\ 149)$ и т. д.

В связи с рассмотрением пифагоровых чисел более или менее естественно возникает вопрос о возможности следующего обобщения задачи: можно ли найти такие целые положительные числа a, b, c, которые удовлетворяли бы уравнению $a^3+b^3=c^3$, или уравнению $a^4+b^4=c^4$, или, вообще говоря, уравнению

$$a^n + b^n = c^n, \tag{3}$$

где показатель n — целое число, большее 2? Ответ был дан Ферма, который пришел к нему умозрительным путем. Ферма изучал одно сочинение Диофанта, известного математика древности, занимавшегося теорией чисел, и имел обыкновение делать примечания на полях книги. Хотя он не затруднял себя тем, чтобы приводить тут же доказательства

многих высказанных им теорем, но все они постепенно в дальнейшем были доказаны — за одним весьма значительным исключением. По поводу пифагоровых чисел Ферма сделал пометку, что *уравнение* (3) *неразрешимо в целых числах, если* $n > 2$, но что найденное им остроумное доказательство этой теоремы слишком длинно, чтобы его можно было поместить на полях книги, с которой он работал.

Это утверждение Ферма в его общей форме никогда и никем впоследствии не было ни доказано, ни опровергнуто, несмотря на усилия целого ряда крупнейших математиков.[1] Правда, теорема была доказана для очень многих значений n, в частности, для всех $n < 619$, но не для всех возможных значений n; вместе с тем не было указано и примера, опровергающего теорему. Хотя сама по себе теорема и не имеет очень большого значения в математическом смысле, но попытки доказать ее положили начало многим важнейшим исследованиям в области теории чисел. Проблема вызвала большой интерес и в более широких кругах — отчасти благодаря премии размером в 100 000 марок, предназначенной для лица, которое впервые даст решение, причем присуждение премии было поручено Геттингенской Академии. Пока послевоенная инфляция в Германии не свела на нет денежную ценность этой премии, ежегодно представлялось громадное число «решений», содержащих ошибки. Даже специалисты-математики не раз были введены в заблуждение и представляли или публиковали доказательства, которые затем отпадали после обнаружения в них иной раз каких-нибудь поверхностных недосмотров. Со времени падения курса марки ажиотаж вокруг проблемы Ферма несколько приутих; и все же пресса не перестает время от времени осведомлять нас о том, что решение найдено каким-нибудь новоявленным «гением».

§ 4. Алгоритм Евклида

1. Общая теория. Читатель прекрасно знаком с обыкновенной процедурой «длинного» деления одного целого числа a на другое число b и знает, что эту процедуру можно продолжать до тех пор, пока остаток не станет меньше, чем делитель. Так, если $a = 648$ и $b = 7$, то мы получаем частное $q = 92$ и остаток $r = 4$.

$$\begin{array}{r|l} 648 & 7 \\ \cline{2-2} \underline{63} & 92 \\ 18 & \\ \underline{14} & \\ 4 & \end{array} \qquad 648 = 7 \cdot 92 + 4.$$

[1] Теорема Ферма была доказана в 1995 г. Подробную историю доказательства этой теоремы можно найти в книге: Сингх С. Великая теорема Ферма. — М.: МЦНМО, 2000. — *Прим. ред. наст. изд.*

По этому поводу можно сформулировать следующую общую теорему: *если a и b — целые числа, причем b отлично от нуля, то можно всегда найти такое целое число q, что*

$$a = b \cdot q + r, \tag{1}$$

где r есть целое число, удовлетворяющее неравенству $0 \leqslant r < b$.

Докажем эту теорему независимо от процедуры длинного деления. Достаточно заметить, что число a или само есть кратное числа b, или же лежит между двумя последовательными кратными b,

$$bq < a < b(q+1) = bq + b.$$

В первом случае равенство (1) оправдывается, причем $r = 0$. Во втором случае из первого неравенства вытекает, что

$$a - bq = r > 0,$$

а из второго — что

$$a - bq = r < b,$$

так что число r в этом случае удовлетворяет условию $0 < r < b$.

Из указанного обстоятельства можно вывести большое число различных важных следствий. Первое из них — это метод для нахождения общего наибольшего делителя двух целых чисел.

Пусть a и b — два каких-то целых числа, не равных одновременно нулю; рассмотрим совокупность всех чисел, на которые делятся и a и b. Эта совокупность, несомненно, конечная, так как если, например, $a \neq 0$, то никакое число, большее чем a, не может быть делителем a. Отсюда следует, что число общих делителей a и b конечно; пусть через d обозначен наибольший из них. Число d называется *общим наибольшим делителем* a и b, и мы условимся обозначать его $d = (a,\, b)$. Так, если $a = 8$, $b = 12$, то непосредственная проверка показывает, что $(8,\, 12) = 4$; если $a = 5$, $b = 9$, то мы точно так же получаем $(5,\, 9) = 1$. Если a и b — достаточно большие числа, например $a = 1804$, $b = 328$, то попытки найти общий наибольший делитель с помощью непосредственных проб довольно утомительны. Короткий и вполне надежный метод вытекает из *алгоритма Евклида.* (Алгоритмом называют всякий систематизированный прием вычисления.) Он основан на том обстоятельстве, что из соотношения вида

$$a = b \cdot q + r \tag{2}$$

необходимо следует, что

$$(a,\, b) = (b,\, r). \tag{3}$$

В самом деле, всякое число u, которое одновременно делит и a и b,

$$a = su, \qquad b = tu,$$

делит также и r, так как $r = a - bq = su - qtu = (s - qt)u$; и обратно, всякое число v, которое одновременно делит b и r,

$$b = s'v, \qquad r = t'v,$$

делит также и a, так как $a = bq + r = s'vq + t'v = (s'q + t')v$. Значит, каждый общий делитель a и b есть вместе с тем общий делитель b и r, и обратно. Но раз совокупность *всех* общих делителей a и b совпадает с совокупностью всех общих делителей b и r, то ясно, что общий *наибольший* делитель a и b должен совпадать с общим *наибольшим* делителем b и r. А это и выражено равенством (3). Мы сейчас убедимся в полезности установленного обстоятельства.

Для этого вернемся к примеру нахождения общего наибольшего делителя чисел 1804 и 328. Обыкновенное «длинное» деление

$$\begin{array}{r|l} 1804 & 328 \\ \cline{2-2} \underline{1640} & 5 \\ 164 & \end{array}$$

приводит нас к заключению, что

$$1804 = 5 \cdot 328 + 164.$$

Отсюда в силу (3) следует, что

$$(1804,\ 328) = (328,\ 164).$$

Заметим, что задача вычисления общего наибольшего делителя (1804, 328) заменена теперь аналогичной задачей, но для меньших чисел. Можно продолжать эту процедуру. Так как

$$\begin{array}{r|l} 328 & 164 \\ \cline{2-2} \underline{328} & 2 \\ 0 & \end{array}$$

то мы получаем дальше $328 = 2 \cdot 164 + 0$, так что $(328,\ 164) = (164,\ 0) = 164$. Значит, $(1804,\ 328) = (328,\ 164) = (164,\ 0) = 164$, и общий наибольший делитель найден.

Эта самая процедура нахождения общего наибольшего делителя двух чисел в геометрической форме описана в «Началах» Евклида. Мы дадим ее общее описание в арифметической форме, исходя из произвольных целых чисел a и b, которые оба одновременно не равны нулю.

Так как сразу ясно, что $(a,\ 0) = a$, то можно допустить, что $b \neq 0$. Последовательные деления приводят нас к цепи равенств

$$\left.\begin{array}{rl} a = bq_1 \ \ + r_1 & (0 < r_1 < b) \\ b = r_1 q_2 + r_2 & (0 < r_2 < r_1) \\ r_1 = r_2 q_3 + r_3 & (0 < r_3 < r_2) \\ r_2 = r_3 q_4 + r_4 & (0 < r_4 < r_3) \\ \dots\dots\dots\dots & \end{array}\right\} \quad (4)$$

Деление продолжается, пока какой-нибудь из остатков $r_1, r_2, r_3, \ldots$ не обратится в нуль. Рассматривая неравенства, выписанные справа, мы

видим, что последовательно получаемые остатки образуют убывающую последовательность положительных чисел:

$$b > r_1 > r_2 > r_3 > r_4 > \ldots > 0. \tag{5}$$

Отсюда ясно, что после конечного числа делений (нужно сделать не более b операций, но часто гораздо меньше, так как разности между соседними r обыкновенно превышают единицу) должен получиться остаток 0:

$$r_{n-2} = r_{n-1}q_n + r_n,$$
$$r_{n-1} = r_n q_{n+1} + 0.$$

Как только это получилось, мы можем утверждать, что

$$(a,\, b) = r_n;$$

другими словами, общий наибольший делитель (a, b) *равен последнему остатку в последовательности* (5). Это следует из многократного применения равенства (3) к соотношениям (4); в самом деле, из этих соотношений следует:

$$(a,\, b) = (b,\, r_1), \qquad (b,\, r_1) = (r_1,\, r_2), \qquad (r_1,\, r_2) = (r_2,\, r_3),$$
$$(r_2,\, r_3) = (r_3,\, r_4), \quad \ldots, \quad (r_{n-1},\, r_n) = (r_n,\, 0) = r_n.$$

Упражнение. Выполните алгоритм Евклида с цепью нахождения общего наибольшего делителя чисел: а) 187, 77; б) 105, 385; в) 245, 193.

Из равенств (4) можно вывести одно чрезвычайно важное свойство общего наибольшего делителя (a, b): *если* $d = (a, b)$, *то можно найти такие целые положительные или отрицательные числа* k *и* l, *что*

$$d = ka + lb. \tag{6}$$

Чтобы убедиться в этом, рассмотрим последовательные остатки (5). Первое из равенств (4) нам дает

$$r_1 = a - q_1 b,$$

так что r_1 может быть записано в форме $k_1 a + l_1 b$ (в данном случае $k_1 = 1$, $l_1 = -q_1$). Из следующего равенства получается

$$r_2 = b - q_2 r_1 = b - q_2(k_1 a + l_1 b) = (-q_2 k_1)a + (1 - q_2 l_1)b = k_2 a + l_2 b.$$

Очевидно, такое же рассуждение можно по очереди применить ко всем остаткам $r_3, r_4, \ldots$, пока мы не придем к представлению

$$r_n = ka + lb,$$

которое и желали получить.

В качестве примера рассмотрим алгоритм Евклида в применении к нахождению (61, 24): общий наибольший делитель есть 1, и интересующее нас представление числа 1 получается из равенств

$$61 = 2 \cdot 24 + 13, \qquad 24 = 1 \cdot 13 + 11, \qquad 13 = 1 \cdot 11 + 2,$$
$$11 = 5 \cdot 2 + 1, \qquad 2 = 2 \cdot 1 + 0.$$

Первое из этих равенств дает

$$13 = 61 - 2 \cdot 24,$$

второе —

$$11 = 24 - 13 = 24 - (61 - 2 \cdot 24) = -61 + 3 \cdot 24,$$

третье —

$$2 = 13 - 11 = (61 - 2 \cdot 24) - (-61 + 3 \cdot 24) = 2 \cdot 61 - 5 \cdot 24$$

и, наконец, четвертое —

$$1 = 11 - 5 \cdot 2 = (-61 + 3 \cdot 24) - 5(2 \cdot 61 - 5 \cdot 24) = -11 \cdot 61 + 28 \cdot 24.$$

2. Применение к основной теореме арифметики. Тот факт, что $d = (a, b)$ всегда может быть записано в форме $d = ka + lb$, позволит нам привести доказательство основной теоремы арифметики, отличное от того, которое было изложено на стр. 42. Сначала в качестве леммы мы докажем следствие, приведенное на стр. 43, а затем уже из него выведем теорему. Таким образом, ход мыслей будет теперь противоположен прежнему.

Лемма. *Если произведение ab делится на простое число p, то или a, или b делится на p.*

Предположим, что a не делится на p; тогда $(a, p) = 1$, так как p имеет лишь два делителя: p и 1. В таком случае можно найти такие целые числа k и l, что

$$1 = ka + lp.$$

Умножая обе части равенства на b, получим:

$$b = kab + lpb.$$

Так как ab делится на p, то можно написать

$$ab = pr,$$

так что

$$b = kpr + lpb = p(kr + lb),$$

и отсюда ясно, что b делится на p. Таким образом, мы установили, что если ab делится на p, но a не делится, то b непременно делится на p; значит, во всяком случае, или a, или b делится на p, раз ab делится на p.

Обобщение на случай произведения трех или большего числа множителей не представляет труда. Например, если abc делится на p, то достаточно дважды применить лемму, чтобы получить заключение, что по меньшей мере один из трех множителей ab и c делится на p. В самом деле, если p не делит ни a, ни b, ни c, то не делит ab и, следовательно, не делит $(ab)c = abc$.

Упражнение. Обобщение этого рассуждения на случай произведения из произвольного числа n множителей требует явного или неявного применения

принципа математической индукции. Воспроизведите все детали соответствующих рассуждений.

Из полученного результата немедленно получается основная теорема арифметики. Предположим, что имеются два разложения целого числа N на простые множители:

$$N = p_1 p_2 \ldots p_r = q_1 q_2 \ldots q_s.$$

Так как p_1 делит левую часть равенства, то должно делить и правую и, значит (см. предыдущее упражнение), должно делить один из множителей q_k. Но q_k — простое число; значит, p_1 должно равняться q_k. Сократив равенство на общий множитель $p_1 = q_k$, обратимся к множителю p_2 и установим *таким же образом, что он равен некоторому* q_t. Сократив на $p_2 = q_t$, переходим, далее, к множителю p_3, и т. д. В конце концов сократятся все множители p, и слева останется 1. Так как q — целые положительные числа, то и справа не может остаться ничего, кроме 1. Итак, числа p и числа q будут попарно равны, независимо от порядка; значит, оба разложения тождественны.

3. Функция Эйлера $\varphi(n)$. Еще раз о теореме Ферма. Говорят, что два целых числа a и b *взаимно простые, если их общий наибольший делитель равен* 1:

$$(a,\, b) = 1.$$

Например, числа 24 и 35 взаимно простые, но числа 12 и 18 не взаимно простые.

Если a и b взаимно простые, то можно подобрать такие целые числа k и l, что

$$ka + lb = 1.$$

Это следует из свойства $(a,\, b)$, отмеченного на стр. 64.

Упражнение. Докажите теорему: *если произведение ab делится на r, причем r и a взаимно простые, то b делится на r.* (*Указание*: если r и a взаимно простые, то можно найти такие целые числа k и l, что

$$kr + la = 1.$$

Затем умножьте обе части равенства на b). Эта теорема обобщает лемму со стр. 65, так как простое число p в том и только в том случае является взаимно простым с a, если a не делится на p.

Пусть n — произвольное целое положительное число; обозначим через $\varphi(n)$ количество таких целых чисел в пределах от 1 до n, которые являются взаимно простыми с числом n. Выражение $\varphi(n)$, впервые введенное Эйлером, представляет собой очень важную теоретико-числовую функцию. Легко подсчитать значения $\varphi(n)$ для нескольких первых зна-

чений n:

$\varphi(1) = 1$,	так	как	1	является взаимно простой с		1
$\varphi(2) = 1$,	»	»	1	»	»	2
$\varphi(3) = 2$,	»	»	1 и 2	являются взаимно простыми с		3
$\varphi(4) = 2$,	»	»	1 и 3	»	»	4
$\varphi(5) = 4$,	»	»	1, 2, 3, 4	»	»	5
$\varphi(6) = 2$,	»	»	1, 5	»	»	6
$\varphi(7) = 6$,	»	»	1, 2, 3, 4, 5, 6	»	»	7
$\varphi(8) = 4$,	»	»	1, 3, 5, 7	»	»	8
$\varphi(9) = 6$,	»	»	1, 2, 4, 5, 7, 8	»	»	9
$\varphi(10) = 4$,	»	»	1, 3, 7, 9	»	»	10

и т. д.

Заметим, что $\varphi(p) = p - 1$, если p — простое число; в самом деле, у числа p нет делителей, кроме 1 и p, и потому все числа $1, 2, \ldots, p-1$ являются взаимно простыми с p. Если n составное и его разложение на простые множители имеет вид

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_r^{\alpha_r},$$

где числа p обозначают различные простые множители, каждый из которых возводится в некоторую степень, то тогда

$$\varphi(n) = n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\ldots\left(1 - \frac{1}{p_r}\right).$$

Например, из разложения $12 = 2^2 \cdot 3$ следует

$$\varphi(12) = 12\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{3}\right) = 12\left(\frac{1}{2}\right)\left(\frac{2}{3}\right) = 4,$$

что легко проверить и непосредственно. Доказательство приведенной теоремы совершенно элементарно, но мы его не приводим.

Упражнение. Пользуясь функцией Эйлера $\varphi(n)$, обобщите теорему Ферма, приведенную на стр. 55. Обобщенная теорема формулируется следующим образом: *если n — целое число и a взаимно просто с n, то*

$$a^{\varphi(n)} \equiv 1 \pmod{n}.$$

4. Непрерывные дроби. Диофантовы уравнения. Алгоритм Евклида, служащий для нахождения общего наибольшего делителя двух целых чисел, сразу же приводит к очень важному методу представления отношения двух целых чисел в виде некоторой сложной дроби особого вида.

Например, в применении к числам 840 и 611 алгоритм Евклида дает ряд равенств

$$840 = 1 \cdot 611 + 229, \qquad 611 = 2 \cdot 229 + 153,$$
$$229 = 1 \cdot 153 + 76, \qquad 153 = 2 \cdot 76 + 1,$$

которые, между прочим, показывают, что $(840, 611) = 1$. Но из этих равенств, с другой стороны, получаются следующие:

$$\frac{840}{611} = 1 + \frac{229}{611} = 1 + \frac{1}{\left(\frac{611}{229}\right)},$$
$$\frac{611}{229} = 2 + \frac{153}{229} = 2 + \frac{1}{\left(\frac{229}{153}\right)},$$
$$\frac{229}{153} = 1 + \frac{76}{153} = 1 + \frac{1}{\left(\frac{153}{76}\right)},$$
$$\frac{153}{76} = 2 + \frac{1}{76}.$$

Комбинируя последние равенства, мы приходим к следующему разложению числа $\frac{840}{611}$:

$$\frac{840}{611} = 1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{76}}}}\ .$$

Выражение вида

$$a = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_n}}}\ , \qquad (7)$$

где все числа a целые положительные, называется *непрерывной дробью.* Алгоритм Евклида дает метод для представления всякого рационального числа в виде такой непрерывной дроби.

Упражнение. Разложите в непрерывные дроби рациональные числа

$$\frac{2}{5}, \quad \frac{43}{30}, \quad \frac{169}{70}.$$

* Непрерывные дроби играют важную роль в той области высшей арифметики, которую иногда называют диофантовым анализом. *Диофантово уравнение* — это алгебраическое уравнение с одним или с несколькими неизвестными, все коэффициенты которого — *целые* числа, причем ставится задача отыскания лишь *целых* его корней. Такое уравнение может или вовсе не иметь решений, или иметь их конечное число, или, наконец, иметь бесконечное множество решений. Простейшее диофантово уравнение — *линейное*, с двумя неизвестными:

$$ax + by = c, \qquad (8)$$

где a, b, c — данные целые числа, и требуется найти целые решения x, y. Полное решение уравнения этого типа может быть найдено посредством алгоритма Евклида.

Прежде всего этот алгоритм позволит нам определить $d = (a, b)$; затем, как мы знаем, при надлежащем выборе целых чисел k и l выполняется равенство

$$ak + bl = d. \qquad (9)$$

Итак, уравнение (8) имеет частное решение $x = k$, $y = l$ в том случае, если $c = d$. Вообще, если c есть кратное d,

$$c = d \cdot q,$$

то из равенства (9) мы выводим

$$a(kq) + b(lq) = dq = c,$$

так что в этом случае уравнение (8) имеет частное решение $x = x^* = kq$, $y = y^* = lq$. Обратно, если уравнение (8) имеет при данном c хоть одно решение x, y, то c должно быть кратным $d = (a, b)$: действительно, d делит и a и b, следовательно, должно делить c. Мы доказали, таким образом, что уравнение (8) имеет (хоть одно) решение в том и только том случае, если c кратно (a, b).

Посмотрим теперь, как, зная одно решение $x = x^*$, $y = y^*$ уравнения (8), определить все прочие решения. Пусть $x = x'$, $y = y'$ есть какое-либо иное решение; тогда $x = x' - x^*$, $y = y' - y^*$ есть решение «однородного» уравнения

$$ax + by = 0. \tag{10}$$

Действительно, из равенств

$$ax' + by' = c \quad \text{и} \quad ax^* + by^* = c$$

посредством вычитания получаем

$$a(x' - x^*) + b(y' - y^*) = 0.$$

Обращаясь теперь к уравнению (10), мы видим, что общее его решение имеет вид $x = \frac{rb}{(a,\, b)}$, $y = -\frac{ra}{(a,\, b)}$ где r — произвольное целое число. (Предоставляем доказательство читателю в качестве упражнения. *Указание*: разделите на (a, b) и воспользуйтесь упражнением на стр. 66.) Затем окончательно будем иметь общее решение уравнения (8):

$$x = x^* + \frac{rb}{(a,\, b)}, \qquad y = y^* - \frac{ra}{(a,\, b)}.$$

Подведем итоги. Линейное диофантово уравнение $ax + by = c$, где a, b и c — целые числа, имеет целые решения в том и только том случае, если c кратно (a, b). В этом случае частное решение $x = x^*$, $y = y^*$ может быть найдено посредством алгоритма Евклида, а самое общее имеет вид

$$x = x^* + \frac{rb}{(a,\, b)}, \qquad y = y^* - \frac{ra}{(a,\, b)},$$

где r — произвольное целое число.

Примеры. Уравнение $3x + 6y = 22$ не имеет целых решений, так как $(3, 6) = 3$ не делит 22.

Уравнение $7x + 11y = 13$ имеет частное решение $x = -39$, $y = 26$, которое находится с помощью следующих вычислений:

$$11 = 1 \cdot 7 + 4, \quad 7 = 1 \cdot 4 + 3, \quad 4 = 1 \cdot 3 + 1, \quad (7,\, 11) = 1,$$
$$1 = 4 - 3 = 4 - (7 - 4) = 2 \cdot 4 - 7 = 2(11 - 7) - 7 = 2 \cdot 11 - 3 \cdot 7.$$

Отсюда следует:

$$7 \cdot (-3) + 11 \cdot (2) = 1,$$
$$7 \cdot (-39) + 11 \cdot (26) = 13.$$

Остальные решения даются формулами

$$x = -39 + 11r, \quad y = 26 - 7r,$$

где r — произвольное целое число.

Упражнение. Решите диофантовы уравнения:

а) $3x - 4y = 29$;

б) $11x + 12y = 58$;

в) $153x - 34y = 51$.

ГЛАВА II

Математическая числовая система

Введение

В дальнейшем мы должны в очень значительной степени расширить понятие числа, связываемое первоначально с натуральным рядом, для того чтобы сконструировать мощный инструмент, способный удовлетворять потребностям и практики, и теории. Исторически — в процессе долгой и неуверенно протекавшей эволюции — нуль, целые отрицательные числа и рациональные дроби приобрели постепенно те же права, что и числа натурального ряда, и в наши дни правилами действий со всеми этими числами прекрасно овладевает средний ребенок школьного возраста. Но для того чтобы обеспечить полную свободу в алгебраических операциях, нужно идти и дальше и охватить расширенным понятием также иррациональные и комплексные числа. Хотя эти обобщения понятия числа употреблялись уже столетия тому назад и на них базируется вся современная математика, но на прочный логический фундамент они были поставлены лишь в недавнее время. В настоящей главе мы дадим очерк основных этапов этого развития.

§ 1. Рациональные числа

1. Рациональные числа как средство измерения. Натуральные числа возникают как абстракция в процессе счета объектов, образующих конечные совокупности. Но в повседневной жизни нам приходится не только *считать* объекты, индивидуально отделенные один от другого, но и *измерять величины*, например такие, как длина, площадь, вес, время. Если мы хотим обеспечить свободу операций с результатами измерения таких величин, могущих неограниченно делиться на части, нам необходимо, не ограничиваясь натуральным рядом, расширить пределы арифметики и создать новый мир чисел. Первый шаг заключается в том, чтобы проблему измерения свести к проблеме счета. Мы выбираем сначала совершенно произвольно *единицу измерения* — фут, ярд, дюйм, фунт, грамм — смотря по случаю, и этой единице приписываем меру 1. Затем мы считаем число таких единиц, входящих в измеряемую величину. Может случиться, что данный кусок свинца весит ровно 54 фунта.

Но в общем случае, как мы замечаем, процесс счета «не сходится»: данная величина не измеряется абсолютно точно выбранной единицей, не оказывается ей кратной. Самое большее, что мы можем сказать в этом случае, — это то, что она заключена между двумя последовательными кратными этой единицы, допустим, между 53 и 54 фунтами. Если так действительно происходит, то мы делаем следующий шаг и вводим новые подъединицы, получающиеся от подразделения первоначальной единицы на некоторое число n равных частей. На обыкновенном языке эти новые подъединицы могут иметь те или иные названия; например, фут подразделяется на 12 дюймов, метр — на 100 сантиметров, фунт — на 16 унций, час — на 60 минут, минута — на 60 секунд, и т. д. Однако в общей математической символике подъединица, получаемая при подразделении первоначальной единицы на n частей, обозначается символом $\frac{1}{n}$, и если рассматриваемая величина содержит ровно m таких подъединиц, то ее мера тогда есть $\frac{m}{n}$. Этот символ называется *дробью* или *отношением* (иногда пишут $m:n$). Последний, и самый существенный, шаг был совершен уже осознанно, после многих столетий накопления отдельных усилий: символ $\frac{m}{n}$ был освобожден от его конкретной связи с процессом измерения и самими измеряемыми величинами и стал рассматриваться как отвлеченное *число*, самостоятельная сущность, уравненная в своих правах с натуральным числом. Если m и n — натуральные числа, то символ $\frac{m}{n}$ называется *рациональным числом.*

Употребление термина «число» (первоначально под «числами» понимали только натуральные числа) применительно к новым символам оправдывается тем обстоятельством, что сложение и умножение этих символов подчиняются тем же законам, что и соответствующие операции над натуральными числами. Чтобы в этом убедиться, нужно сначала определить, в чем заключаются сложение и умножение рациональных чисел, а также определить, какие рациональные числа признаются равными между собой. Эти определения, как всем известно, таковы:

$$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd}, \qquad \frac{a}{b}\cdot\frac{c}{d}=\frac{ac}{bd}, \qquad \frac{a}{a}=1, \qquad \frac{ac}{bc}=\frac{a}{b}, \tag{1}$$

где a, b, c, d — произвольные натуральные числа. Например,

$$\frac{2}{3}+\frac{4}{5}=\frac{2\cdot 5+3\cdot 4}{3\cdot 5}=\frac{10+12}{15}=\frac{22}{15}, \qquad \frac{2}{3}\cdot\frac{4}{5}=\frac{2\cdot 4}{3\cdot 5}=\frac{8}{15},$$

$$\frac{3}{3}=1, \qquad \frac{8}{12}=\frac{2\cdot 4}{3\cdot 4}=\frac{2}{3}.$$

Эти самые определения мы *вынуждены* принять, если имеем в виду использовать рациональные числа для измерения длин, площадей и т. п. Но с более строгой логической точки зрения эти правила сложения и умножения и это толкование равенства по отношению ко вновь вводимым символам устанавливаются независимо по определению, не будучи

обусловлены какой-либо иной необходимостью, кроме взаимной совместимости (непротиворечивости) и пригодности к практическим приложениям. Исходя из определений (1), можно показать, что *основные законы арифметики натуральных чисел продолжают сохраняться и в области всех рациональных чисел*:

$$\left.\begin{array}{ll} p+q=q+p & \text{(коммутативный закон сложения)},\\ p+(q+r)=(p+q)+r & \text{(ассоциативный закон сложения)},\\ pq=qp & \text{(коммутативный закон умножения)},\\ p(qr)=(pq)r & \text{(ассоциативный закон умножения)},\\ p(q+r)=pq+pr & \text{(дистрибутивный закон)}. \end{array}\right\} \quad (2)$$

Так, например, доказательство коммутативного закона сложения в случае дробей ясно из следующих равенств:

$$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd}=\frac{cb+da}{db}=\frac{c}{d}+\frac{a}{b};$$

здесь первое и последнее равенства оправдываются определением сложения (1), а среднее есть следствие коммутативных законов сложения и умножения в области натуральных чисел. Читатель сможет, если пожелает, проверить таким же образом четыре остальных закона.

2. Возникновение надобности в рациональных числах внутри самой математики. Принцип обобщения. Независимо от «практического» основания для введения рациональных чисел существует основание более глубокое и носящее в известном смысле еще более принудительный характер. Эту сторону дела мы рассмотрим здесь совершенно независимо от приведенных выше рассуждений. В обычной арифметике натуральных чисел мы всегда можем выполнять основные *прямые* операции — сложение и умножение. Но *обратные* операции — вычитание и деление — не всегда выполнимы. Разность $b-a$ двух натуральных чисел a и b есть по определению такое натуральное число c, что $a+c=b$, т. е. это есть решение уравнения $a+x=b$. Но в области натуральных чисел символ $b-a$ имеет смысл лишь при ограничении $b>a$, так как только при этом условии уравнение $a+x=b$ имеет решением натуральное число. На пути к снятию этого ограничения серьезный шаг был сделан уже тогда, когда был введен символ 0 для обозначения $a-a$. Но еще более значительным успехом было введение символов -1, -2, -3, ... и вместе с тем определения

$$(b-a)=-(a-b)$$

для случая $b<a$: после этого можно было утверждать, что и вычитание обладает свойством неограниченной выполнимости *в области всех целых — положительных и отрицательных — чисел*. Вводя новые сим-

волы $-1, -2, -3, \ldots$ и тем самым расширяя числовую область, мы обязаны, конечно, *определить операции* со вновь вводимыми числами таким образом, чтобы *первоначальные правила арифметических операций не были нарушены.* Так, например, правило

$$(-1)\cdot(-1)=1, \tag{3}$$

которое лежит в основе умножения отрицательных чисел, есть следствие нашего желания сохранить дистрибутивный закон $a(b+c)=ab+ac$. Действительно, если бы мы, скажем, декларировали, что $(-1)\cdot(-1)=-1$, то, полагая $a=-1$, $b=1$, $c=-1$, получили бы $(-1)\cdot(1-1)=-1-1=-2$, тогда как на самом деле $(-1)\cdot(1-1)=(-1)\cdot 0=0$.

Понадобилось немало времени, чтобы среди математиков было хорошо осознано, что «правило знаков» (3) и вместе с ним все прочие определения, относящиеся как к отрицательным числам, так и к дробям, никак не могут быть «доказаны». Они *создаются*, или декларируются, нами самими с целью обеспечить свободу операций и притом без нарушения основных арифметических законов. Что может — и должно — быть доказываемо, так это только то, что если эти определения приняты, то тем самым сохранены основные законы арифметики: коммутативный, ассоциативный и дистрибутивный. Даже великий Эйлер пользовался совершенно неубедительной аргументацией, желая показать, что $(-1)\cdot(-1)$ «должно» равняться $+1$. Он говорил: «Рассматриваемое произведение может быть только или $+1$, или -1; но -1 быть не может, так как $-1=(+1)\cdot(-1)$.»

Совершенно подобно тому, как введение отрицательных целых чисел и нуля расчищает путь для неограниченной выполнимости вычитания, введение дробных чисел устраняет арифметические препятствия, мешающие выполнять деление. Отношение, или частное, $x=\frac{b}{a}$ двух целых чисел определяется как решение уравнения

$$ax=b \tag{4}$$

и существует как *целое число* только в том случае, если a есть делитель b. Но если это не так (например, при $a=2$, $b=3$), то мы просто вводим новый символ $\frac{b}{a}$, называемый дробью и подчиненный условию, выражающемуся равенством $a\cdot\frac{b}{a}=b$, так что $\frac{b}{a}$ есть решение (4) «по определению». Изобретение дробей как новых числовых символов обеспечивает неограниченную выполнимость деления, *за исключением деления на нуль*, которое *исключается раз навсегда.*

Выражения вроде $\frac{1}{0}$, $\frac{3}{0}$, $\frac{0}{0}$ и т. п. останутся для нас символами, лишенными смысла. Если бы мы допустили деление на 0, то из верного равенства $0\cdot 1=0\cdot 2$ вывели бы неверное следствие $1=2$. Иногда

бывает целесообразно обозначать такие выражения символом «бесконечность», *однако с условием, чтобы не делалось даже попытки оперировать этим символом так, как будто бы он подчинялся обычным законам арифметики.*

Теперь нам ясны принципы, согласно которым сконструирована система *всех рациональных чисел* — целых и дробных, положительных и отрицательных. В этой расширенной области не только полностью оправдываются формальные законы — ассоциативный, коммутативный и дистрибутивный, — но и уравнения $a + x = b$ и $ax = b$ всегда имеют решения $x = b - a$ и $x = \frac{b}{a}$ с единственной оговоркой, что в случае второго уравнения a не должно равняться нулю. Иными словами, в области рациональных чисел так называемые *рациональные операции* — сложение, вычитание, умножение и деление — выполнимы неограниченно и не выводят за пределы области. Такие замкнутые числовые области называются *полями.* Мы повстречаемся с дальнейшими примерами полей ниже, в этой же главе, а также в главе III.

Расширение области посредством введения новых символов, совершаемое таким образом, что законы, которые имели место в первоначальной области, сохраняются и в расширенной, является типичным примером характерного для математики *принципа обобщения.* Переход путем обобщения от натуральных чисел к рациональным удовлетворяет одновременно и теоретической потребности в снятии ограничений, которые наложены на вычитание и деление, и вместе с тем — практической потребности в числах, пригодных для фиксации результатов измерений. Именно тот факт, что рациональные числа идут навстречу сразу теоретической и практической потребностям, придает им особую важность. Как мы видели, расширение понятия числа совершилось путем введения новых абстрактных символов вроде 0, -2 или $\frac{3}{4}$. В наше время мы оперируем этими символами бегло и уверенно, не вдумываясь в их природу, и трудно даже себе представить, что еще в XVII столетии они пользовались доверием гораздо в меньшей степени, чем натуральные числа, что ими если и пользовались, то с известным сомнением и трепетом. Свойственное человеческому сознанию стремление цепляться за «конкретное» — воплощаемое в ряде натуральных чисел — обусловливает ту медленность, с которой протекала неизбежная эволюция. Логически безупречная арифметическая система может быть сконструирована не иначе, как в отвлечении от действительности.

3. Геометрическое представление рациональных чисел. Выразительное геометрическое представление системы рациональных чисел может быть получено следующим образом.

На некоторой прямой линии, «числовой оси», отметим отрезок от 0 до 1 (рис. 8). Тем самым устанавливается длина единичного отрезка, которая, вообще говоря, может быть выбрана произвольно. Положительные и отрицательные целые числа тогда изображаются совокупностью равноотстоящих точек на числовой оси, именно, положительные числа отмечаются вправо, а отрицательные — влево от точки 0. Чтобы изобразить числа со знаменателем n, разделим каждый из полученных отрезков единичной длины на n равных частей; точки деления будут изображать дроби со знаменателем n. Если сделать так для значений n, соответствующих всем натуральным числам, то каждое рациональное число будет изображено некоторой точкой числовой оси. Эти точки мы условимся называть «рациональными»; вообще, термины «рациональное число» и «рациональная точка» будем употреблять как синонимы.

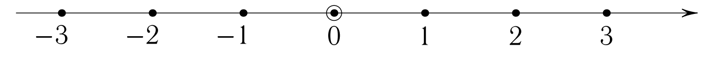


Рис. 8. Числовая ось

В главе I, § 1 было определено соотношение неравенства $A < B$ для натуральных чисел. На числовой оси это соотношение отражено следующим образом: если натуральное число A меньше, чем натуральное число B, то точка A лежит левее точки B. Так как указанное геометрическое соотношение устанавливается для *любой пары* рациональных точек, то естественно пытаться обобщить арифметическое отношение неравенства таким образом, чтобы сохранить этот геометрический порядок для рассматриваемых точек. Это удается, если принять следующее определение: говорят, что рациональное число A *меньше*, чем рациональное число B ($A < B$), или что число B *больше*, чем число A ($B > A$), если разность $B - A$ положительна. Отсюда следует (при $A < B$), что точки (числа) *между* A и B — это те, которые одновременно $> A$ и $< B$. Каждая такая пара точек A и B, вместе со всеми точками между ними, называется *сегментом* (или *отрезком*) и обозначается $[A, B]$ (а множество одних только промежуточных точек — *интервалом* (или *промежутком*), обозначаемым (A, B)).

Расстояние произвольной точки A от начала 0, рассматриваемое как положительное число, называется *абсолютной величиной* A и обозначается символом

$$|A|.$$

Понятие «абсолютная величина» определяется следующим образом: если $A \geqslant 0$, то $|A| = A$; если $A < 0$, то $|A| = -A$. Ясно, что если числа A и B имеют один и тот же знак, то справедливо равенство $|A + B| = |A| + |B|$; если же A и B имеют разные знаки, то $|A + B| < |A| + |B|$. Соединяя эти

два результата вместе, мы приходим к общему неравенству

$$|A+B| \leqslant |A| + |B|,$$

которое справедливо независимо от знаков A и B.

Факт фундаментальной важности выражается следующим предложением: *рациональные точки расположены на числовой прямой всюду плотно.* Смысл этого утверждения тот, что внутри всякого интервала, как бы он ни был мал, содержатся рациональные точки. Чтобы убедиться в справедливости высказанного утверждения, достаточно взять число n настолько большое, что интервал $\left(0, \dfrac{1}{n}\right)$ будет меньше, чем данный интервал (A, B); тогда по меньшей мере одна из точек вида $\dfrac{m}{n}$ окажется внутри данного интервала. Итак, не существует такого интервала на числовой оси (даже самого маленького, какой только можно вообразить), внутри которого не было бы рациональных точек. Отсюда вытекает дальнейшее следствие: во всяком интервале содержится бесконечное множество рациональных точек. Действительно, если бы в некотором интервале содержалось лишь конечное число рациональных точек, то внутри интервала, образованного двумя соседними такими точками, рациональных точек уже не было бы, а это противоречит тому, что только что было доказано.

§ 2. Несоизмеримые отрезки. Иррациональные числа, пределы

1. Введение. Если мы станем сравнивать по величине два прямолинейных отрезка a и b, то не исключена возможность, что a содержится в b в точности целое число раз r. В таком случае длина отрезка b очень просто выражается через длину отрезка a: длина b в r раз больше, чем длина a. Может случиться и так, что целого числа r, которое обладало бы указанным свойством, не существует; но при этом возможно, что, разделив отрезок a на некоторое число, скажем n, равных частей $\left(\text{каждая длины } \dfrac{a}{n}\right)$ и взяв целое число m таких частей, мы в точности получим отрезок b:

$$b = \frac{m}{n}\,a. \tag{1}$$

Если осуществляется соотношение вида (1), то говорят, что два отрезка a и b *соизмеримы*, так как они обладают некоторой «общей мерой»: таковой является отрезок длины $\dfrac{a}{n}$, который содержится в отрезке a ровно n раз, а в отрезке b ровно m раз. Некоторый отрезок b соизмерим или несоизмерим с отрезком a в зависимости от того, можно или нельзя подобрать два таких натуральных числа m и n $(n \neq 0)$, что имеет

место равенство (1). Обращаясь к рис. 9, предположим, что в качестве отрезка a избран единичный отрезок $[0, 1]$, и рассмотрим всевозможные отрезки, у которых один из концов совпадает с 0. Тогда из этих отрезков те и только те будут соизмеримы с единичным отрезком, у которых второй конец совпадает с некоторой рациональной точкой $\frac{m}{n}$.

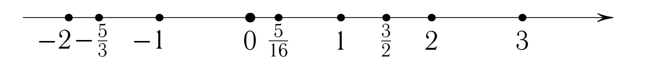


Рис. 9. Рациональные точки

Для практической цели измерения рациональных чисел всегда совершенно достаточно. Даже с точки зрения теоретической, поскольку рациональные точки расположены всюду плотно, могло бы показаться, что все точки на числовой оси — рациональные. Если бы дело обстояло именно так, то всякий отрезок был бы соизмерим с единичным. Но дело обстоит не так просто, и в установлении этого обстоятельства заключается одно из самых поразительных открытий в математике: оно было сделано уже в древнейшие времена (в школе Пифагора). *Существуют несоизмеримые отрезки*, или иначе (если мы допустим, что каждому отрезку соответствует некоторое число, выражающее его длину), *существуют иррациональные числа*. Осознание этого факта было научным событием величайшей значимости, почти откровением. Весьма возможно, что именно оно положило начало тому, что мы теперь считаем строгим математическим методом и рассматриваем как вклад в науку, сделанный древними греческими математиками. Без сомнения, это замечательное открытие глубоко повлияло на всю математику и даже философию от древних времен и до наших дней.

Евдоксова теория несоизмеримых величин, изложенная в геометрической форме в «Началах» Евклида, представляет собой тончайшее достижение греческой математики (ее изложение обыкновенно пропускается в разжиженных пересказах Евклида, предназначенных для школьного обучения). Эта теория получила подобающую ей высокую оценку лишь в конце XIX столетия — после того как усилиями Дедекинда, Кантора и Вейерштрасса была создана строгая теория иррациональных чисел. Мы изложим в дальнейшем эту теорию в ее современном арифметическом аспекте.

Прежде всего установим: *диагональ квадрата несоизмерима с его стороной.* Предположим, что сторона квадрата избрана в качестве единицы длины, длину же диагонали обозначим через x. Тогда, согласно теореме Пифагора, мы получаем:

$$x^2 = 1^2 + 1^2 = 2.$$

(Такое число x обозначают символом $\sqrt{2}$.) Если бы x было соизмеримо с единицей, то можно было бы найти два таких целых числа p и q,

что $x = \frac{p}{q}$, и тогда мы пришли бы к равенству

$$p^2 = 2q^2. \tag{2}$$

Можно допустить, что дробь $\frac{p}{q}$ несократима, иначе мы с самого начала сократили бы ее на общий наибольший делитель чисел p и q. С правой стороны имеется 2 в качестве множителя, и потому p^2 есть четное число, и, значит, само p — также четное, так как квадрат нечетного числа есть нечетное число. В таком случае можно положить $p = 2r$. Тогда равенство (2) принимает вид:

$$4r^2 = 2q^2, \quad \text{или} \quad 2r^2 = q^2.$$

Так как с левой стороны теперь имеется 2 в качестве множителя, значит, q^2, а следовательно, и q — четное. Итак, и p и q — четные числа, т. е. делятся на 2, а это противоречит допущению, что дробь $\frac{p}{q}$ несократима. Итак, равенство (2) невозможно, и x не может быть рациональным числом.

Иначе этот результат можно сформулировать, утверждая, что $\sqrt{2}$ есть число иррациональное.

Только что приведенное рассуждение показывает, что иной раз самое простейшее геометрическое построение приводит к отрезку, несоизмеримому с единицей. Если такой отрезок будет отложен с помощью циркуля на числовой оси от точки 0, то построенная таким образом точка (конец отрезка) не совпадает ни с какой рациональной точкой. Итак, система *рациональных точек* (хотя и всюду плотная) *не покрывает всей числовой оси*. Наивному сознанию, несомненно, может показаться странным и парадоксальным, что всюду плотное множество рациональных точек не покрывает всей прямой. Никакая наша «интуиция» не поможет нам «увидеть» иррациональные точки или отличить их от рациональных. Нет ничего удивительного в том, что открытие несоизмеримого потрясло греческих математиков и мыслителей и что его существование и в наши дни продолжает производить впечатление на людей, склонных к углубленным размышлениям.

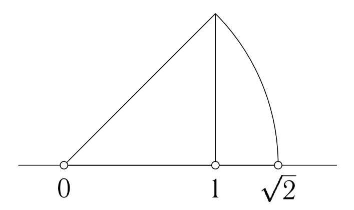


Рис. 10. Построение числа $\sqrt{2}$

Не представило бы труда сконструировать столько отрезков, несоизмеримых с единицей, сколько бы мы пожелали. Концы всех таких отрезков — при условии, что их начала совпадают с точкой 0,— образуют совокупность иррациональных точек. Заметим теперь, что нашим руководящим принципом уже при введении рациональных дробей было желание *обеспечить возможность измерения длин отрезков посредством чисел*, и тот же принцип продолжает руководить нами и тогда,

когда речь идет о несоизмеримых отрезках. Если мы требуем, чтобы существовало взаимное соответствие между числами, с одной стороны, и точками на прямой линии — с другой, то неизбежно приходится ввести в рассмотрение *иррациональные числа.*

Подводя итоги до сих пор сказанному, мы констатируем, что иррациональное число обозначает длину отрезка, несоизмеримого с единицей. В следующих разделах мы должны будем уточнить это несколько смутное и всецело геометрическое определение и в результате придем к определению, более удовлетворительному с точки зрения логической строгости. Рассматривая этот вопрос, мы будем вначале исходить из десятичных дробей.

Упражнения. 1) Докажите, что числа $\sqrt[3]{2}$, $\sqrt{3}$, $\sqrt{5}$, $\sqrt[3]{3}$ иррациональные. (*Указание*: воспользуйтесь леммой на стр. 65.)

2) Докажите, что числа $\sqrt{2}+\sqrt{3}$ и $\sqrt{2}+\sqrt[3]{2}$ иррациональные. (*Указание*: если бы, например, первое из этих чисел было рациональным числом r, то, написав $\sqrt{3}=r-\sqrt{2}$ и возведя в квадрат, мы заключили бы, что $\sqrt{2}$ есть рациональное число.)

3) Докажите, что число $\sqrt{2}+\sqrt{3}+\sqrt{5}$ иррациональное. Попробуйте придумать еще подобные и более общие примеры.

2. Десятичные дроби: конечные и бесконечные. Чтобы покрыть числовую ось везде плотным множеством точек, нет необходимости использовать *всю* совокупность рациональных чисел: достаточно, например, ограничиться только теми числами, которые возникают при подразделении единичного отрезка на 10, потом на 100, 1000 и т. д. равных частей. Получающиеся при этом точки деления соответствуют «десятичным дробям». Так, числу $0{,}12=\frac{1}{10}+\frac{2}{100}$ соответствует точка, расположенная в первом единичном интервале, во втором «подынтервале» длины 10^{-1}, и именно она есть начальная точка третьего «подподынтервала» длины 10^{-2} $\left(a^{-n}\text{ означает }\frac{1}{a^n}\right)$. Если такого рода *десятичная дробь* содержит n знаков после запятой, то она имеет вид

$$f=z+a_1\cdot 10^{-1}+a_2\cdot 10^{-2}+a_3\cdot 10^{-3}+\ldots+a_n\cdot 10^{-n},$$

где z — целое число, а коэффициенты a — цифры 0, 1, 2, …, 9, обозначающие число десятых, сотых и т. д. Сокращенно число f записывается в десятичной системе следующим образом: $z{,}a_1a_2a_3\ldots a_n$. Мы убеждаемся непосредственно, что такого рода десятичные дроби могут представлены виде обыкновенных дробей $\frac{p}{q}$, где $q=10^n$; так, например,

$$f=1{,}314=1+\frac{3}{10}+\frac{1}{100}+\frac{4}{1000}=\frac{1314}{1000}.$$

Если окажется, что p и q имеют общий множитель, то дробь можно сократить, и тогда знаменатель будет некоторым делителем числа 10^n.

С другой стороны, несократимая дробь, у которой знаменатель не есть делитель некоторой степени 10, не может быть представлена в виде десятичной дроби указанного типа. Например, $\frac{1}{5} = \frac{2}{10} = 0{,}2$; $\frac{1}{250} = \frac{4}{1000} = 0{,}004$; но $\frac{1}{3}$ не может быть написана как десятичная дробь с конечным числом n десятичных знаков, как бы ни было велико n: в самом деле, из равенства вида

$$\frac{1}{3} = \frac{b}{10^n}$$

следовало бы

$$10^n = 3b,$$

а последнее равенство невозможно, так как 3 не входит множителем ни в какую степень числа 10.

Возьмем теперь на числовой оси какую-нибудь точку P, которая не соответствует никакой конечной десятичной дроби; можно, например, взять рациональную точку $\frac{1}{3}$ или иррациональную точку $\sqrt{2}$. Тогда в процессе последовательного подразделения единичного интервала на 10 равных частей точка P никогда не окажется в числе точек деления: она будет находиться внутри десятичных интервалов, длина которых будет неограниченно уменьшаться; концы этих интервалов соответствуют конечным десятичным дробям и приближают точку P с какой угодно степенью точности. Рассмотрим несколько подробнее этот процесс приближения.

Предположим, что точка P лежит в первом единичном интервале. Сделаем подразделение этого интервала на 10 равных частей, каждая длины 10^{-1}, и предположим, что точка P попадает, скажем, в третий из этих интервалов. На этой стадии мы можем утверждать, что P заключена между десятичными дробями 0,2 и 0,3. Подразделяем снова интервал от 0,2 до 0,3 на 10 равных частей, каждая длины 10^{-2}, и обнаружим, что P попадает, допустим, в четвертый из этих интервалов. Подразделяя его, как раньше, видим, что точка P попадает в первый интервал длины 10^{-3}. Теперь можно сказать, что точка P заключена между 0,230 и 0,231. Этот процесс может быть продолжен до бесконечности и приводит к бесконечной последовательности цифр a_1, a_2, a_3, $\ldots$, a_n, $\ldots$, обладающей таким свойством: каково бы ни было n, точка P заключена в интервале I_n, у которого начальная точка есть $0{,}a_1a_2a_3\ldots a_{n-1}a_n$, а конечная — $0{,}a_1a_2a_3\ldots a_{n-1}(a_n+1)$, причем длина I_n равна 10^{-n}. Если станем полагать по порядку $n = 1, 2, 3, 4, \ldots$, то увидим, что каждый из интервалов I_1, I_2, I_3, $\ldots$ содержится в предыдущем, причем их длины 10^{-1}, 10^{-2}, 10^{-3}, $\ldots$ неограниченно уменьшаются. Мы скажем, более кратко, что точка P заключена в *стягивающуюся последовательность десятичных интервалов*. Например, если точка P есть $\frac{1}{3}$, то все

цифры a_1, a_2, a_3, ... равны 3, и P заключена в любом интервале I_n от 0,333...33 до 0,333...34, т. е. $\frac{1}{3}$ больше чем 0,333...33 и меньше чем 0,333...34, сколько бы ни взять цифр после запятой. Мы скажем в этих обстоятельствах, что n-значная десятичная дробь 0,333...33 «стремится к $\frac{1}{3}$», когда число цифр n неограниченно возрастает. И мы условимся писать

$$\frac{1}{3} = 0{,}333\ldots,$$

причем точки обозначают, что десятичная дробь может быть продлена «до бесконечности».

Иррациональная точка $\sqrt{2}$, которая была рассмотрена в пункте 1, также приводит к бесконечной десятичной дроби. Но закон, которому подчиняются последовательные цифры десятичного разложения, на этот раз далеко не очевиден. Мы затрудняемся указать формулу, которая давала бы цифру, стоящую на n-м месте, хотя можно вычислить столько цифр, сколько мы пожелали бы себе заранее назначить:

$$1^2 = 1 < 2 < 2^2 = 4$$
$$(1{,}4)^2 = 1{,}96 < 2 < (1{,}5)^2 = 2{,}25$$
$$(1{,}41)^2 = 1{,}9881 < 2 < (1{,}42)^2 = 2{,}0264$$
$$(1{,}414)^2 = 1{,}999396 < 2 < (1{,}415)^2 = 2{,}002225$$
$$(1{,}4142)^2 = 1{,}99996164 < 2 < (1{,}4143)^2 = 2{,}00024449 \quad \text{и т. д.}$$

В качестве общего определения мы скажем, что точка P, которая не может быть представлена в виде десятичной дроби с конечным числом десятичных знаков, представляется в виде *бесконечной десятичной дроби* $z{,}a_1a_2a_3\ldots$, если, каково бы ни было n, точка P лежит в интервале длины 10^{-n} с начальной точкой $z{,}a_1a_2a_3\ldots a_n$.

Таким образом, мы устанавливаем соответствие между всеми точками числовой оси и всеми (конечными или бесконечными) десятичными дробями. Теперь мы попытаемся ввести предварительное определение: «число» есть *конечная или бесконечная десятичная дробь.* Те бесконечные десятичные дроби, которые не представляют рационального числа, называются *иррациональными числами.* До середины XIX столетия соображения, подобные приведенным выше, казались достаточными для объяснения того, как устроена система рациональных и иррациональных чисел — *числовой континуум.* Необычайные успехи математики, достигнутые начиная с XVII столетия, в частности, развитие аналитической геометрии и дифференциального и интегрального исчислений, твердо базировались именно на таком представлении о системе чисел. Однако в период критического пересмотра принципов и консолидации результатов стало ощущаться все более и более явственно, что понятие иррационального числа должно быть подвергнуто

более точному и глубокому анализу. Но, прежде чем перейти к очерку современной теории числового континуума, нам придется рассмотреть и разобрать — на более или менее интуитивной основе — одно из математических понятий капитальной значимости — понятие *предела.*

Упражнение. Вычислите приближенно $\sqrt[3]{2}$ и $\sqrt[3]{5}$ с ошибкой, не превышающей 10^{-2}.

3. Пределы. Бесконечные геометрические прогрессии. Как мы видели в предыдущем пункте, иногда случается, что некоторое рациональное число s приближается последовательностью других рациональных чисел s_n, причем индекс n принимает последовательно все значения 1, 2, 3, ... Так, например, можно взять: $s = \frac{1}{3}$, тогда $s_1 = 0{,}3$, $s_2 = 0{,}33$, $s_3 = 0{,}333$ и т. д. Вот еще пример. Разобьем единичный интервал на две равные части, вторую половину — снова на две равные части, вторую из полученных двух частей — снова на две равные части и т. д., пока наименьший из полученных таким образом интервалов не станет равным 2^{-n}, где n — сколь угодно большое наперед заданное число, например, $n = 100$, $n = 100\,000$ и т. д. Затем, складывая вместе все интервалы, кроме самого последнего, мы получаем общую длину

$$s_n = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots + \frac{1}{2^n}. \tag{3}$$

Легко понять, что s_n отличается от 1 на $\left(\frac{1}{2}\right)^n$ и что эта разность становится сколь угодно малой, или «стремится к нулю», при неограниченном возрастании n. Говорить, что эта разность *равна* нулю, когда n равно «бесконечности», не имеет никакого смысла. Бесконечное в математике связывается с некоторым *процессом*, не имеющим конца, и никогда не связывается с актуальной *величиной.* Желая описать поведение s_n, мы говорим, что *сумма s_n стремится к пределу* 1, *когда n стремится к бесконечности*, и пишем

$$1 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots, \tag{4}$$

причем то, что возникает справа, есть *бесконечный ряд.* Последнее «равенство» не следует понимать в том смысле, что имеется в виду сложить вместе бесконечное число слагаемых: это только *сокращенная* запись того факта, что 1 есть предел конечных сумм s_n, получающийся, когда n *стремится* к бесконечности (и ни в коем случае *не равно* бесконечности). Итак, равенство (4), заканчивающееся неопределенным символом «$+\ldots$», как бы стенографирует некоторую точную мысль, выражаемую, по неизбежности, длинным рядом слов:

«1 равна пределу (при n, стремящемся к бесконечности) выражения

$$s_n = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots + \frac{1}{2^n}\text{»}. \tag{5}$$

Еще более кратко и более выразительно пишут следующим образом:

$$s_n \to 1 \quad \text{при} \quad n \to \infty. \tag{6}$$

Говоря о пределах, рассмотрим еще пример. Пусть перед нами имеется бесконечная последовательность различных степеней числа q:

$$q, \ \ q^2, \ \ q^3, \ \ q^4, \ \ \ldots, \ \ q^n, \ \ \ldots$$

Если $-1 < q < 1$, например, $q = \frac{1}{3}$ или $q = -\frac{4}{5}$, то q^n стремится к нулю при неограниченном возрастании n. При этом если q — отрицательное число, то знаки q^n чередуются: за $+$ следует $-$, и обратно; таким образом, q^n стремится к нулю «с двух сторон». Так, если $q = \frac{1}{3}$, то $q^2 = \frac{1}{9}$, $q^3 = \frac{1}{27}$, $q^4 = \frac{1}{81}$, $\ldots$; но если $q = -\frac{1}{2}$, то $q^2 = \frac{1}{4}$, $q^3 = -\frac{1}{8}$, $q^4 = \frac{1}{16}$, $\ldots$ Мы утверждаем, что *предел* q^n, *когда* n *стремится к бесконечности, равен нулю*, или, символически,

$$q^n \to 0 \quad \text{при} \quad n \to \infty, \quad \text{ если } -1 < q < 1. \tag{7}$$

(Между прочим, если $q > 1$ или $q < -1$, то q^n уже не стремится к нулю, а неограниченно возрастает по абсолютной величине.)

Приведем строгое доказательство утверждения (7). Мы видели на стр. 34, что при любом целом положительном значении n и при условии $p > -1$ имеет место неравенство $(1+p)^n \geqslant 1 + np$. Пусть q — какое-то положительное число, меньшее единицы, например, $q = \frac{9}{10}$. Тогда можно положить $q = \frac{1}{1+p}$, где $p > 0$. Отсюда следует

$$\frac{1}{q^n} = (1+p)^n \geqslant 1 + np > np,$$

или же (см. определение (4) на стр. 74)

$$0 < q^n < \frac{1}{p} \cdot \frac{1}{n}.$$

Значит, q^n заключено между постоянным числом 0 и числом $\frac{1}{p} \cdot \frac{1}{n}$, которое стремится к нулю при неограниченном возрастании n (так как p — постоянное). После этого ясно, что $q^n \to 0$. Если q — отрицательное число, то мы положим $q = -\frac{1}{1+p}$, и тогда q^n будет заключено между числами $-\frac{1}{p} \cdot \frac{1}{n}$ и $\frac{1}{p} \cdot \frac{1}{n}$; рассуждение заканчивается так же, как раньше.

Рассмотрим теперь *геометрическую прогрессию*

$$s_n = 1 + q + q^2 + q^3 + \ldots + q^n. \tag{8}$$

(Частный случай $q = \frac{1}{2}$ был рассмотрен выше.) Как уже было показано (см. стр. 32), сумма s_n может быть представлена в более простой и

сжатой форме. Умножая s_n на q, мы получаем

$$qs_n = q + q^2 + q^3 + q^4 + \ldots + q^{n+1} \tag{8a}$$

и, вычитая (8a) из (8), убеждаемся, что все члены, кроме 1 и q^{n+1}, взаимно уничтожаются. В результате будем иметь

$$(1-q)s_n = 1 - q^{n+1},$$

или же, деля на $1-q$,

$$s_n = \frac{1-q^{n+1}}{1-q} = \frac{1}{1-q} - \frac{q^{n+1}}{1-q}.$$

С понятием предела мы встретимся, если заставим n неограниченно возрастать. Мы видели только что, что $q^{n+1} = q \cdot q^n$ стремится к нулю, если $-1 < q < 1$, и отсюда можем заключить:

$$s_n \to \frac{1}{1-q} \quad \text{при} \quad n \to \infty, \quad \text{если } -1 < q < 1. \tag{9}$$

Тот же результат можно записать, пользуясь *бесконечным рядом*

$$1 + q + q^2 + q^3 + \ldots = \frac{1}{1-q}, \quad \text{если } -1 < q < 1. \tag{10}$$

Например,

$$1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots = \frac{1}{1-\frac{1}{2}} = 2$$

в полном соответствии с равенством (4); подобным же образом

$$\frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3} + \frac{9}{10^4} + \ldots = \frac{9}{10} \cdot \frac{1}{1-\frac{1}{10}} = 1,$$

или, иначе, $0{,}9999\ldots = 1$. Совершенно так же конечная дробь 0,2374 и бесконечная дробь 0,23739999… представляют одно и то же число.

В главе VI мы вернемся к общему обсуждению понятия предела, рассматривая вопрос с современной, логически более строгой точки зрения.

Упражнения. 1) Докажите, что

$$1 - q + q^2 - q^3 + q^4 - \ldots = \frac{1}{1+q}, \quad \text{если } |q| < 1.$$

2) Каков предел последовательности $a_1, a_2, a_3, \ldots$, где $a_n = \frac{n}{n+1}$? (*Указание*: напишите данное выражение $\frac{n}{n+1}$ в виде $1 - \frac{1}{n+1}$ и обратите внимание на то, что вычитаемое стремится к нулю.)

3) Каков предел $\frac{n^2+n+1}{n^2-n+1}$ при $n \to \infty$? (*Указание*: напишите это выражение в виде

$$\frac{1 + \frac{1}{n} + \frac{1}{n^2}}{1 - \frac{1}{n} + \frac{1}{n^2}}.$$

4) Предполагая q по абсолютной величине меньшим чем 1, докажите, что $1 + 2q + 3q^2 + 4q^3 + \ldots = \frac{1}{(1-q)^2}$. (*Указание*: воспользуйтесь результатом упражнения 3 на стр. 36.)

5) Каков предел бесконечного ряда

$$1 - 2q + 3q^2 - 4q^3 + \ldots \text{ ?}$$

6) Вычислите пределы выражений

$$\frac{1+2+3+\ldots+n}{n^2}, \quad \frac{1^2+2^2+3^2+\ldots+n^2}{n^3}, \quad \frac{1^3+2^3+3^3+\ldots+n^3}{n^4}.$$

(*Указание*: воспользуйтесь результатами, полученными на стр. 31–33.)

4. Рациональные числа и периодические десятичные дроби. Такие рациональные числа $\frac{p}{q}$, которые не могут быть представлены в виде конечных десятичных дробей, разлагаются в бесконечные десятичные дроби посредством обыкновенного приема «длинного» деления. На каждой ступени этого процесса возникает остаток, не равный нулю, иначе дробь оказалась бы конечной. Различные возникающие остатки могут быть только целыми числами от 1 до $q - 1$, так что имеется всего $q - 1$ возможностей для значений этих остатков. Это значит, что после q делений некоторый остаток k появится во второй раз. Но тогда все следующие остатки также будут повторяться в том же порядке, в каком они уже появлялись после первого возникновения остатка k. Таким образом, *десятичное разложение всякого рационального числа обладает свойством периодичности*; после некоторого числа десятичных знаков одна и та же группа десятичных знаков начинает повторяться бесконечное число раз. Например, $\frac{1}{6} = 0{,}166666666\ldots$; $\frac{1}{7} = 0{,}142857142857142857\ldots$; $\frac{1}{11} = 0{,}09090909\ldots$; $\frac{122}{1100} = 0{,}1109090909\ldots$; $\frac{11}{90} = 0{,}122222222\ldots$ и т. д. (Заметим по поводу тех рациональных чисел, которые представляются в виде конечной десятичной дроби, что у этой конечной дроби можно вообразить после последнего ее десятичного знака бесконечно повторяющуюся цифру 0, и, таким образом, рассматриваемые рациональные числа не исключаются из данной выше общей формулировки.) Из приведенных примеров видно, что у некоторых из десятичных разложений, соответствующих рациональным числам, периодическому «хвосту» предшествует непериодическая «голова».

Обратно, можно показать, что *все периодические дроби представляют собой рациональные числа.* Рассмотрим, например, бесконечную периодическую дробь

$$p = 0{,}3322222\ldots$$

Можно написать: $p = \frac{33}{100} + 10^{-3} \cdot 2(1 + 10^{-1} + 10^{-2} + \ldots)$. Выражение в

скобках есть бесконечная геометрическая прогрессия:

$$1 + 10^{-1} + 10^{-2} + 10^{-3} + \ldots = \frac{1}{1 - \frac{1}{10}} = \frac{10}{9}.$$

Значит,

$$p = \frac{33}{100} + 10^{-3} \cdot 2 \cdot \frac{10}{9} = \frac{2970 + 20}{9 \cdot 10^3} = \frac{2990}{9000} = \frac{299}{900}.$$

В общем случае доказательство строится таким же образом, но затруднено необходимостью вводить несколько громоздкие обозначения. Рассмотрим периодическую дробь общего вида

$$p = 0{,}a_1a_2a_3 \ldots a_m b_1b_2b_3 \ldots b_n b_1b_2b_3 \ldots b_n \ldots$$

Обозначим через $B = 0{,}b_1b_2b_3 \ldots b_n$ периодическую часть нашего разложения. Тогда можно написать

$$p = 0{,}a_1a_2a_3 \ldots a_m + 10^{-m}B(1 + 10^{-n} + 10^{-2n} + 10^{-3n} + \ldots).$$

Выражение в скобках — бесконечная геометрическая прогрессия, для которой $q = 10^{-n}$. Сумма этой прогрессии, согласно формуле (10) предыдущего пункта, равна $\frac{1}{1 - 10^{-n}}$, и потому

$$p = 0{,}a_1a_2a_3 \ldots a_m + \frac{10^{-m} \cdot B}{1 - 10^{-n}}.$$

Упражнения. 1) Разложите в десятичные дроби следующие рациональные числа: $\frac{1}{11}$, $\frac{1}{13}$, $\frac{2}{13}$, $\frac{3}{13}$, $\frac{1}{17}$, $\frac{2}{17}$, и определите периоды разложений.

*2) Число 142 857 обладает тем свойством, что при умножении его на 2, 3, 4, 5 или 6 в нем совершаются только перестановки цифр. Объясните это свойство, исходя из разложения числа $\frac{1}{7}$ в десятичную дробь.

3) Разложите числа, приведенные в упражнении 1, в бесконечные дроби с основаниями 5, 7 и 12.

4) Разложите число $\frac{1}{3}$ в двоичную дробь.

5) Напишите разложение 0,11212121... Установите, какое число оно представляет при основаниях 3 или 5.

5. Общее определение иррациональных чисел посредством стягивающихся отрезков. На стр. 82 мы ввели предварительное определение: «число» есть конечная или бесконечная десятичная дробь. Мы условились вместе с тем десятичные дроби, не представляющие рационального числа, называть иррациональными числами. На основе результатов, полученных в предыдущем пункте, мы можем теперь предложить следующую формулировку: «*числовой континуум, или система действительных чисел* («действительные» числа противопоставляются здесь «мнимым», или «комплексным», см. § 5), *есть совокупность всевозможных бесконечных десятичных дробей*». (Приписывая нули,

можно, как уже было отмечено, конечную десятичную дробь написать в виде бесконечной, или есть другой способ: последнюю цифру дроби a заменить на $a - 1$ и к ней приписать бесчисленное множество девяток. Так, мы видели, например, что $0{,}999\ldots = 1$, — см. п. 3.)

Рациональные числа суть периодические дроби; иррациональные числа суть непериодические дроби. Но и такое определение не представляется вполне удовлетворительным: действительно, мы видели в главе I, что самой природой вещей десятичная система ничем особым не выделяется из других возможных; таким же образом можно было бы оперировать, например, двоичной системой. По этой причине является чрезвычайно желательным дать более общее определение числового континуума, независимое от специального выбора основания 10 или любого иного. Вероятно, простейший метод для введения такого обобщения заключается в следующем.

Рассмотрим на числовой оси некоторую последовательность I_1, I_2, I_3, ..., I_n, ... отрезков с рациональными концами; предположим, что каждый следующий отрезок содержится в предыдущем и что длина n-го отрезка I_n стремится к нулю при неограниченном возрастании n. Такую последовательность «вложенных» друг в друга отрезков мы будем называть *последовательностью стягивающихся отрезков.* В случае десятичных отрезков длина I_n равна 10^{-n}, но с таким же успехом она могла бы равняться, скажем, 2^{-n}, или можно ограничиться хотя бы тем требованием, чтобы она была меньше $\frac{1}{n}$. Дадим теперь следующую формулировку, которую будем рассматривать как основной геометрический постулат: *какова бы ни была последовательность стягивающихся отрезков, существует одна и только одна точка числовой оси, которая одновременно содержится во всех отрезках.* (Совершенно ясно, что существует не более одной такой точки, так как длины отрезков стремятся к нулю, а две различные точки не могли бы содержаться в отрезке, длина которого была бы меньше, чем расстояние между точками.) Эта точка, по определению, и называется *действительным числом*; если она не является рациональной, то называется *иррациональным числом.* С помощью такого определения мы устанавливаем полное соответствие между точками и числами. Здесь не прибавлено ничего существенно нового: всего лишь определению числа как бесконечной десятичной дроби придана более общая форма.

Все же читателя в этом месте могут охватить известные сомнения, которые следует признать вполне обоснованными. Что же на самом деле представляет собой та «точка» на числовой оси, которая, как мы допускаем, содержится одновременно во всех стягивающихся отрезках последовательности в случае, если она не соответствует рациональному числу? Наш ответ таков: существование на числовой оси

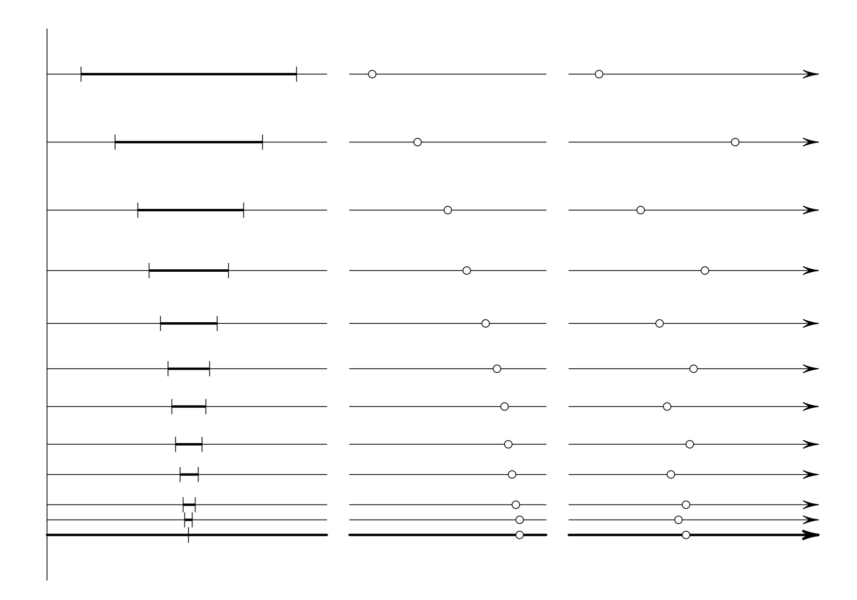


Рис. 11. Стягивающиеся отрезки. Пределы последовательностей

(рассматриваемой как геометрический образ) точки, содержащейся во всех стягивающихся отрезках с рациональными концами, есть основной *геометрический постулат*. Нет надобности делать редукцию, приводя его к иным математическим предложениям. Мы принимаем его, как принимаем в математике другие аксиомы или постулаты, основываясь на его интуитивной правдоподобности и на его полезности, обнаруживающейся при построении логически последовательной системы математических предложений. Чисто формально мы могли бы исходить из числовой прямой, которую мыслили бы как совокупность одних только рациональных точек, и затем определили бы иррациональную точку как *символ, обозначающий некоторую последовательность стягивающихся отрезков*. Иррациональная точка полностью определяется последовательностью стягивающихся рациональных отрезков, длины которых стремятся к нулю. Значит, наш основной постулат на самом деле способен служить определением. Принять такое определение, после того как мы были приведены к последовательности стягивающихся отрезков интуитивным ощущением, утверждающим «существование» иррациональной точки, — значит отбросить «костыли интуиции», на которые опиралось наше рассуждение, и осознать, что все *математические свойства* иррациональных точек могут быть понимаемы и представляемы как свойства последовательностей стягивающихся отрезков.

С чисто математической точки зрения в данном случае важно то обстоятельство, что, приняв определение иррационального числа как

последовательности стягивающихся отрезков, мы приобретаем возможность дать определения сложения, умножения и т. д., а также отношений неравенства, являющихся непосредственным обобщением соответствующих определений в поле рациональных чисел, и притом с сохранением всех основных законов, действующих в поле рациональных чисел. Так, например, чтобы определить сумму двух иррациональных чисел α и β исходя из двух последовательностей стягивающихся отрезков, определяющих числа α и β, построим новую последовательность стягивающихся отрезков, складывая соответственно начальные и конечные точки отрезков, входящих в состав данных последовательностей. То же можно сделать с произведением $\alpha\beta$, разностью $\alpha - \beta$ и частным α/β. И можно показать на основе этих определений, что арифметические законы, рассмотренные в § 1 этой главы, при переходе к иррациональным числам не нарушаются. Подробности, сюда относящиеся, мы опускаем.

Проверка всех этих законов проста и производится непосредственно без особых затруднений, но могла бы показаться несколько скучноватой начинающему читателю, который, естественно, интересуется скорее тем, что можно сделать с помощью математики, чем анализом ее логических основ. Нередко случается, что новейшие учебники математики отталкивают читателя именно тем, что с первых же страниц дают педантическое обоснование системы действительных чисел. Читатель, спокойно игнорирующий эти страницы, пусть успокоит свою совесть сознанием того факта, что вплоть до конца XIX столетия все великие математики делали свои открытия на основе «наивной» концепции числового континуума, доставляемой непосредственно интуицией.

Наконец, с физической точки зрения, определение иррационального числа посредством последовательности стягивающихся отрезков естественно уподобляется определению числового значения некоторой доступной наблюдению величины — путем ряда измерений, производимых последовательно со все возрастающей точностью. Всякая операция, совершаемая, скажем, с целью определения длины некоторого отрезка, практически осмыслена лишь в пределах некоторой возможной погрешности, величину которой определяет точность инструмента. Так как рациональные числа расположены на прямой всюду плотно, то никакая физическая операция, как бы точна она ни была, не позволит различить, является ли данная длина рациональной или же иррациональной. Таким образом, могло бы показаться, что в иррациональных числах нет никакой необходимости для адекватного описания физических явлений. Но, как мы увидим в главе VI, при математическом описании физических явлений истинное преимущество, приобретаемое посредством привлечения иррациональных чисел, заключается в чрезвычайном упрощении этого описания — именно благодаря свободному использованию понятия предела, основой которого является числовой континуум.

***6. Иные методы определения иррациональных чисел. Дедекиндовы сечения.** Несколько иной путь для определения иррациональных чисел был избран Рихардом Дедекиндом (1831–1916), одним из самых выдающихся основоположников логического и философского анализа основ математики. Его статьи — «Stetigkeit und irrationale Zahlen»[1] (1872) и «Was sind und was sollen die Zahlen?»[2] (1887) — оказали глубокое влияние на исследование основных принципов математики. Дедекинд предпочитал общие абстрактные концепции конкретным построениям вроде последовательностей стягивающихся отрезков. Его процедура базируется на идее «сечения»; мы сейчас опишем, что это такое.

Предположим, что каким-то способом удалось разбить совокупность всех рациональных чисел на два класса A и B таким образом, что всякое число b класса B больше, чем всякое число a класса A. Всякое разбиение такого рода называется *сечением* в области рациональных чисел. Если произведено сечение, то должна осуществиться одна из следующих трех логически мыслимых возможностей.

1) *Существует наибольший элемент a^* в классе A.* Такое положение вещей имеет место, например, в том случае, если к классу A отнесены все рациональные числа $\leqslant 1$, к классу B — все рациональные числа > 1.

2) *Существует наименьший элемент b^* в классе B.* Это происходит, например, в том случае, если к классу A отнесены все рациональные числа < 1, к классу B — все рациональные числа $\geqslant 1$.

3) *Нет ни наибольшего элемента в классе A, ни наименьшего в классе B.* Сечение этого рода получится, например, в том случае, если к классу A отнесены все рациональные числа, квадрат которых меньше чем 2, а к классу B — все рациональные числа, квадрат которых больше чем 2. Классами A и B исчерпываются все рациональные числа, так как было показано, что такого рационального числа, квадрат которого равен 2, не существует.

Такой случай, когда в классе A есть наибольший элемент a^* и вместе с тем в классе B — наименьший элемент b^*, логически немыслим, так как тогда рациональное число $\dfrac{a^*+b^*}{2}$, заключенное как раз между a^* и b^*, было бы больше, чем наибольший элемент в A, и меньше, чем наименьший элемент в B, и, значит, не могло бы принадлежать ни к A, ни к B.

В третьем случае, когда нет ни наибольшего рационального числа в классе A, ни наименьшего в классе B, тогда, по Дедекинду, сечение определяет, или, лучше, *представляет собой*, некоторое иррациональ-

[1] «Непрерывность и иррациональные числа». — *Прим. ред.*
[2] «Что такое числа и чем они должны быть?» — *Прим. ред.*

ное число. Не составит труда проверить, что определение Дедекинда согласуется с определением, в основе которого находятся вложенные отрезки: из всякой последовательности вложенных отрезков I_1, I_2, I_3, ... мы получаем сечение, если отнесем к классу A все те рациональные числа, которые меньше, чем левый конец хотя бы одного интервала I_n, к классу B — все прочие рациональные числа.

В философском отношении определение иррациональных чисел по Дедекинду находится на более высоком уровне абстракции, так как оно не ограничивает ни в чем того математического закона, который определяет классы A и B. Другой, более конкретный метод для определения континуума действительных чисел принадлежит Георгу Кантору (1845–1918). На первый взгляд резко отличный как от метода вложенных отрезков, так и от метода сечений, он, однако, эквивалентен любому из них в том смысле, что числовой континуум, получающийся на основе всех трех методов, обладает одними и теми же свойствами. Идея Кантора базируется на тех обстоятельствах, что 1) действительные числа можно трактовать как бесконечные десятичные дроби, 2) бесконечные десятичные дроби можно рассматривать как пределы конечных десятичных дробей. Чтобы не связывать себя зависимостью от десятичных дробей, мы, следуя Кантору, принимаем, что всякая «сходящаяся» последовательность рациональных чисел $a_1, a_2, a_3, \ldots$ определяет действительное число. При этом «сходимость» понимается в том смысле, что разность $(a_m - a_n)$ между двумя членами последовательности стремится к нулю, если m и n одновременно и независимо друг от друга неограниченно возрастают. (Как раз последовательные десятичные приближения обладают этим свойством: любые два из них после n-го отличаются меньше чем на 10^{-n}.) Так как одно и то же действительное число по методу Кантора может быть определяемо самыми разнообразными последовательностями рациональных чисел, то приходится добавить, что две последовательности $a_1, a_2, a_3, \ldots$ и $b_1, b_2, b_3, \ldots$ определяют одно и то же действительное число, если разность $a_n - b_n$ стремится к нулю при неограниченном возрастании n. Идя по пути, намеченному Кантором, нетрудно определить сложение и т. д.

§ 3. Замечания из области аналитической геометрии[1]

1. Основной принцип. Уже начиная с XVII в. числовой континуум, принимаемый как нечто само собой разумеющееся или же подвергаемый более или менее поверхностному критическому анализу, стал основой математики, в частности, аналитической геометрии и дифференциального и интегрального исчислений.

Введение числового континуума дает возможность сопоставить

[1] Читателю, не вполне освоившемуся с предметом этого параграфа, рекомендуется обратиться к упражнениям, которые помещены в приложении в конце книги, стр. 513 и дальше.

каждому отрезку прямой в качестве его «длины» некоторое определенное действительное число. Но можно пойти и дальше. Не только длина, но и *всякий вообще геометрический объект, всякая геометрическая операция могут найти свое место в царстве чисел.* Решительные шаги в направлении арифметизации геометрии был сделаны не позднее 1629 г. Ферма (1601–1665) и в 1637 г. Декартом (1596–1650). Основная идея аналитической геометрии заключается в использовании «координат» — чисел, связанных (координированных) с данным *геометрическим* объектом и полностью этот объект характеризующих. Большинству читателей известны так называемые прямоугольные, или декартовы, координаты, служащие для того, чтобы фиксировать положение произвольной точки на плоскости. Мы исходим из двух неподвижных взаимно перпендикулярных прямых на плоскости, «оси x» и «оси y», и к ним относим каждую точку. Эти оси рассматриваются как ориентированные числовые прямые, причем измерение совершается с помощью одного и того же единичного отрезка. Каждой точке P (рис. 12) сопоставлены две координаты x и y. Они получаются следующим образом. Рассмотрим ориентированный отрезок (вектор), идущий из «начала» O в точку P, и затем спроектируем ортогонально этот вектор на обе оси, получая ориентированный отрезок OP' на оси x и такой же отрезок OQ' на оси y. Два числа x и y, измеряющие соответственно ориентированную длину отрезков OP' и OQ', называются *координатами* точки P. Обратно, если x и y — два произвольных наперед заданных числа, то соответствующая точка P определяется однозначно. Если числа x и y оба положительные, то P попадает в *первый квадрант* координатной системы (рис. 13); если оба отрицательные, то в третий; если x положительно, а y отрицательно, то в четвертый, и, наконец, если x отрицательно, а y положительно, то во второй.

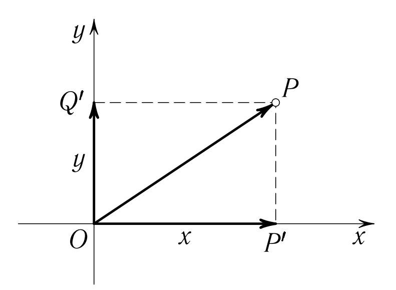


Рис. 12. Прямоугольные координаты точки

Расстояние между точкой P_1 с координатами x_1, y_1 и точкой P_2 с координатами x_2, y_2 дается формулой

$$d^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2. \qquad (1)$$

Это немедленно следует из пифагоровой теоремы (рис. 14).

2. Уравнения прямых и кривых линий. Если C есть неподвижная точка с координатами $x = a$, $y = b$, то геометрическое место всех точек P, находящихся от точки C на данном расстоянии r, есть окруж-

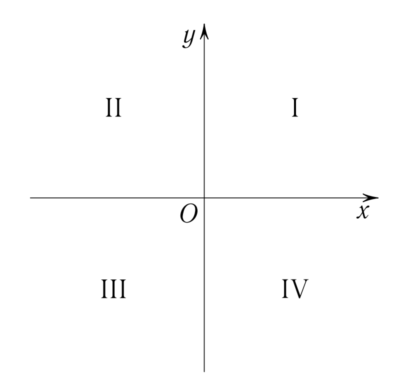


Рис. 13. Четыре квадранта

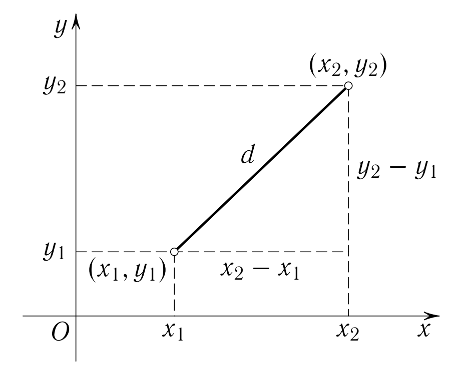


Рис. 14. Расстояние между двумя точками

ность с центром C и радиусом r. Из формулы для расстояния между двумя точками (1) следует, что точки этой окружности имеют координаты x, y, удовлетворяющие уравнению

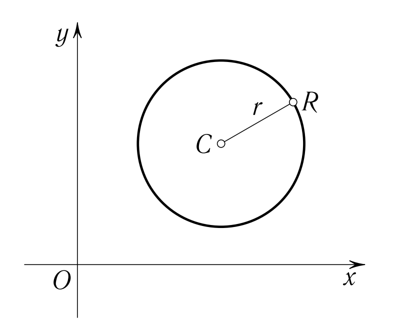


Рис. 15. Окружность

$$(x-a)^2 + (y-b)^2 = r^2. \qquad (2)$$

Это уравнение называется *уравнением окружности*, так как оно выражает полное (необходимое и достаточное) условие того, что точка P с координатами x, y лежит на окружности с центром C и радиусом r. Если скобки раскрыть, уравнение принимает вид

$$x^2 + y^2 - 2ax - 2by = k, \qquad (3)$$

где $k = r^2 - a^2 - b^2$. Обратно, если задано уравнение вида (3), причем a, b и k — произвольные постоянные и сумма $k + a^2 + b^2$ положительна, то с помощью алгебраической процедуры «дополнения до квадрата» мы можем написать то же уравнение в форме

$$(x-a)^2 + (y-b)^2 = r^2,$$

где $r^2 = k + a^2 + b^2$. И тогда ясно, что уравнение (3) определяет окружность радиуса r, центр которой — в точке C с координатами a, b.

Уравнение прямой линий еще проще по своей форме. Так, например, уравнение оси x имеет вид $y = 0$, так как координата y равна нулю для всех точек этой оси и ни для каких иных точек. Точно так же ось y имеет уравнение $x = 0$. Прямые, проходящие через начало и делящие пополам углы между осями, имеют уравнения $x = y$ и $x = -y$. Легко показать, что всякая прямая линия имеет уравнение вида

$$ax + by = c, \qquad (4)$$

где a, b, c — постоянные, характеризующие эту прямую. Как и в других случаях, смысл уравнения (4) тот, что пары действительных чисел x и y, удовлетворяющих этому уравнению, являются координатами некоторой точки на прямой, и обратно.

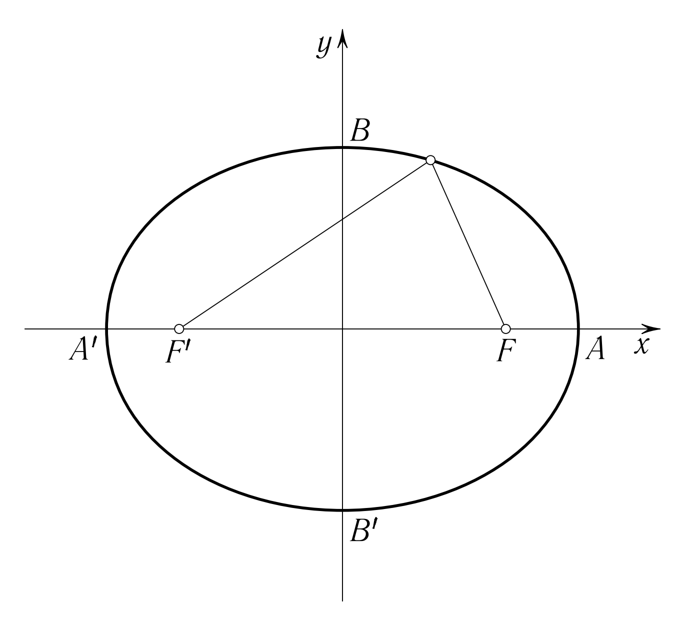


Рис. 16. Эллипс с фокусами

Может быть, читатель учил в школе, что уравнение вида

$$\frac{x^2}{p^2} + \frac{y^2}{q^2} = 1 \tag{5}$$

представляет *эллипс* (рис. 16). Эта кривая пересекает ось x в точках $A(p,\,0)$ и $A'(-p,\,0)$ и ось y в точках $B(0,\,q)$ и $B'(0,\,-q)$. (Обозначение $P(x,\,y)$ или, еще короче, $(x,\,y)$, вводится ради краткости и должно быть расшифровано так: «точка P с координатами x и y».) Если $p > q$, то отрезок AA' длины $2p$ называется *большой осью* эллипса, а отрезок BB' длины $2q$ — его *малой осью.* Эллипс есть геометрическое место точек P, сумма расстояний которых от точек $F(\sqrt{p^2 - q^2},\,0)$ и $F'(-\sqrt{p^2 - q^2},\,0)$ равна $2p$. Читатель сможет проверить это в качестве упражнения, применяя формулу (1). Точки F и F' называются *фокусами* эллипса, а отношение $e = \dfrac{\sqrt{p^2 - q^2}}{p}$ называется его *эксцентриситетом.*

Уравнение вида

$$\frac{x^2}{p^2} - \frac{y^2}{q^2} = 1 \tag{6}$$

представляет *гиперболу.* Эта кривая состоит из двух ветвей, пересе-

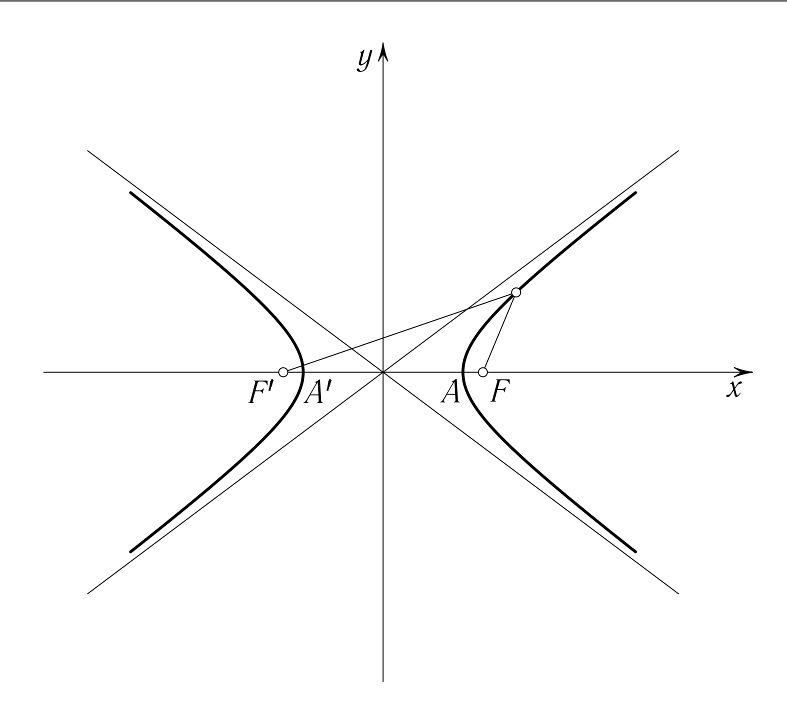


Рис. 17. Гипербола с фокусами

кающих ось x соответственно в точках $A(p, 0)$ и $A'(-p, 0)$ (рис. 17). Отрезок AA' длины $2p$ называется *«действительной» осью* гиперболы. Гипербола, удаляясь в бесконечность, приближается к двум прямым $qx \pm py = 0$, но так с ними и не пересекается; эти прямые называются *асимптотами* гиперболы. Гипербола есть геометрическое место точек P, разность расстояний которых до двух точек $F(\sqrt{p^2+q^2}, 0)$ и $F'(-\sqrt{p^2+q^2}, 0)$ по абсолютной величине равна $2p$. Эти точки в случае гиперболы тоже называются *фокусами*; под *эксцентриситетом* гиперболы понимают отношение $e = \dfrac{\sqrt{p^2+q^2}}{p}$.

Уравнение

$$xy = 1 \tag{7}$$

также определяет гиперболу, но такую, для которой асимптотами являются две оси (рис. 18). Уравнение этой «равносторонней» гиперболы геометрически означает, что площадь прямоугольника $OP'PQ'$ (см. рис. 12), связанного с точкой P, для всякой точки P кривой равна 1. Равносторонняя гипербола несколько более общего вида

$$xy = c, \tag{7a}$$

где c — постоянная, представляет собой частный случай гиперболы в

том же смысле, в каком окружность представляет собой частный случай эллипса. Отличительная характеристика равносторонней гиперболы заключается в том, что ее две асимптоты (в нашем случае — две оси) взаимно перпендикулярны.

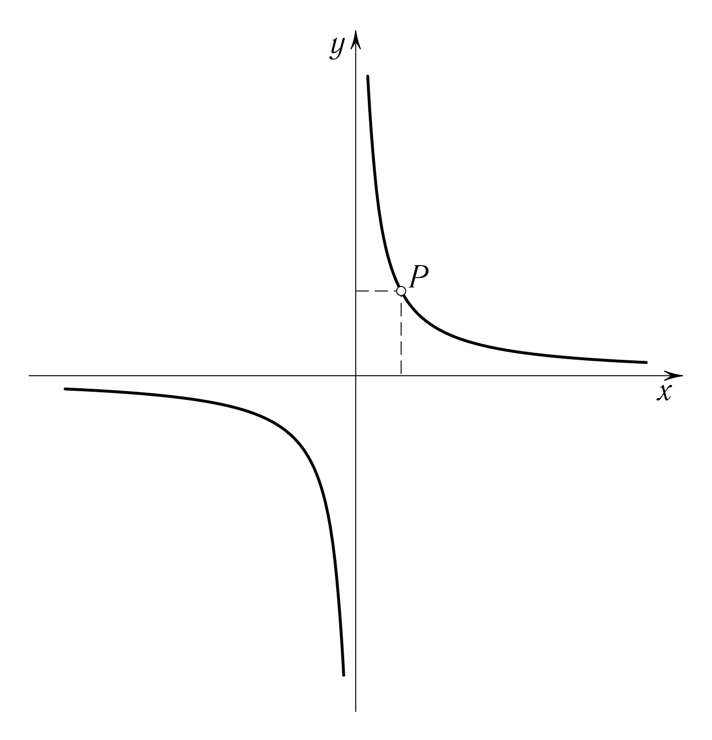


Рис. 18. Равносторонняя гипербола. Площадь прямоугольника, определенного точкой $P(x,\, y)$, равна 1

Во всем этом для нас самым интересным является руководящая идея: геометрические объекты могут полностью описываться в арифметической или алгебраической форме. То же справедливо и относительно геометрических операций. Например, если нам требуется найти точки пересечения двух прямых, то мы рассматриваем два их уравнения

$$\begin{aligned} ax + by &= c, \\ a'x + b'y &= c', \end{aligned} \tag{8}$$

и для нахождения общей точки этих двух прямых достаточно решить систему (8); решение дает нам координаты искомой точки. Таким же образом точки пересечения двух произвольных кривых (скажем, окружности $x^2 + y^2 - 2ax - 2by = k$ и прямой $ax + by = c$) находятся посредством совместного решения их уравнений.

§ 4. Математический анализ бесконечного

1. Основные понятия. Последовательность натуральных чисел

$$1,\ 2,\ 3, \ldots$$

представляет собой первый и самый важный пример бесконечного множества. Не нужно видеть ничего таинственного в том, что она — бесконечная, что у нее «нет конца»: как бы велико ни было натуральное число n, можно построить другое, следующее за ним число, еще большее — $n + 1$. Но при переходе от прилагательного «бесконечный», означающего просто-напросто «не имеющий конца», к существительному «бесконечность» никоим образом не следует привносить допущения, что «бесконечность», обыкновенно изображаемая особым символом ∞, может быть рассматриваема как обыкновенное число. Нельзя включить символ ∞ в числовую систему действительных чисел, не нарушая при этом основных законов арифметики. И тем не менее идея бесконечности пронизывает всю математику, так как математические объекты изучаются обыкновенно не как индивидуумы — каждый в отдельности, а как члены классов или совокупностей, содержащих бесчисленное множество элементов одного и того же типа; таковы совокупности натуральных чисел, действительных чисел или же треугольников на плоскости. Именно по этой причине возникает необходимость в точном математическом анализе бесконечного. Современная теория множеств, созданная Георгом Кантором и его школой в конце XIX столетия, приступив к разрешению этой задачи, достигла значительных успехов. Канторова теория множеств глубоко проникла во многие области математики и оказала на них огромное влияние; она стала играть особо выдающуюся роль в исследованиях, связанных с логическим и философским обоснованием математики. Исходным в канторовой теории является общее понятие совокупности или множества. При этом имеется в виду собрание объектов (элементов), которое определяется некоторым правилом, позволяющим с полной определенностью судить о том, входит ли данный объект в число элементов собрания или не входит. Примерами могут служить множество всех натуральных чисел, множество всех периодических десятичных дробей, множество всех действительных чисел или множество всех прямых в трехмерном пространстве.

Для того чтобы сравнивать множества с точки зрения «количества» содержащихся в них элементов, нужно ввести основное в этой теории понятие «эквивалентности» множеств. Если элементы двух множеств A и B могут быть приведены в попарное соответствие такого рода, что каждому элементу множества A сопоставлен один и только один элемент множества B, а каждому элементу множества B сопоставлен один и только один элемент множества A, то установленное таким образом соответствие называется *взаимно однозначным*, а о самих множествах A и B

тогда говорят, что они между собой *эквивалентны*. Понятие эквивалентности в случае конечных множеств совпадает с обыкновенным понятием *числового равенства*, так как два конечных множества в том и только том случае могут быть приведены во взаимно однозначное соответствие, если содержат одно и то же число элементов. На этом и основывается, нужно заметить, идея счета: когда мы «считаем» элементы множества, то процесс счета как раз и заключается в установлении взаимно однозначного соответствия между элементами множества и числами $1, 2, \ldots, n$.

Чтобы установить эквивалентность двух конечных множеств, иногда нет необходимости «считать» элементы. Так, например, не считая, можно утверждать, что конечное множество кругов единичного радиуса эквивалентно множеству их центров.

Перенося понятие эквивалентности на бесконечные множества, Кантор имел в виду создать «арифметику» бесконечного. Множество действительных чисел и множество точек на прямой линии эквивалентны, так как после того, как выбраны начало и единичный отрезок, данная прямая становится «числовой прямой», и каждой ее точке P в качестве координаты взаимно однозначно сопоставляется некоторое совершенно определенное действительное число x:

$$P \leftrightarrow x.$$

Четные числа образуют правильное подмножество множества всех *натуральных чисел*, а *все целые числа* образуют правильное подмножество множества *всех рациональных чисел*. (Говоря о «правильном» подмножестве некоторого множества S, мы имеем в виду множество S', состоящее из элементов множества S, но *не из всех* его элементов.) Совершенно ясно, что *если данное множество конечно*, т. е. содержит какое-то число n элементов и не более того, *то оно не может быть эквивалентно никакому своему правильному подмножеству*, так как всякое правильное его подмножество содержало бы самое большее $n-1$ элемент. Но *если данное множество содержит бесконечное число элементов*, то, как это ни парадоксально, *оно может быть эквивалентно некоторому своему правильному подмножеству*. Например, схема

$$\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 \ldots & n \ldots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ 2 & 4 & 6 & 8 & 10 \ldots & 2n \ldots \end{array}$$

устанавливает взаимно однозначное соответствие между множеством натуральных чисел и множеством всех четных целых положительных чисел, и эти два множества оказываются эквивалентными, хотя второе есть правильное подмножество первого. Такое противоречие с ходячей истиной «целое больше своей части» показывает, какие сюрпризы нас ждут в области «арифметики бесконечного».

2. Счетность множества рациональных чисел и несчетность континуума. Одно из первых открытий Кантора в области анализа бесконечного заключалось в том, что *множество рациональных чисел* (содержащее в качестве правильного подмножества бесконечное множество натуральных чисел и потому само бесконечное) эквивалентно множеству *натуральных чисел*. На первый взгляд кажется странным, что всюду плотное множество рациональных чисел не более богато элементами, чем множество натуральных чисел, элементы которого «рассеяны» редко и стоят на значительном расстоянии один от другого. И в самом деле, с *сохранением порядка возрастания* нельзя расположить положительные рациональные числа так, как это можно сделать с натуральными: самое маленькое число a будет первым, следующее за ним по величине b вторым, и т. д.; дело в том, что рациональные числа расположены везде плотно, и потому ни для одного из них нельзя указать «следующего по величине». Но Кантор заметил, что если отказаться от требования «располагать по величине», то тогда оказывается возможным расставить все рациональные числа в ряд r_1, r_2, r_3, r_4, $\ldots$, подобный ряду натуральных чисел. Такое расположение предметов некоторого множества в виде последовательности часто называют *пересчетом* («нумерацией») этого множества. Множества, для которых пересчет может быть выполнен, называются *счетными* или *исчислимыми*. Указывая один из способов пересчета множества рациональных чисел и устанавливая, таким образом, его счетность, Кантор тем самым показал, что это множество эквивалентно множеству натуральных чисел, так как схема

$$\begin{array}{ccccc} 1 & 2 & 3 & 4 \ldots & n \ldots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ r_1 & r_2 & r_3 & r_4 \ldots & r_n \ldots \end{array}$$

создает взаимно однозначное соответствие между двумя множествами. Мы опишем сейчас один из возможных способов пересчета множества рациональных чисел.

Каждое рациональное число записывается в виде $\frac{a}{b}$, где a и b — целые числа; все эти числа могут быть расположены в виде таблицы, где число $\frac{a}{b}$ стоит в a-м столбце и в b-й строчке. Например, $\frac{3}{4}$ станет в третьем столбце и в четвертой строчке таблицы. Предположим, что все свободные места, или «клеточки», в таблице заполнены соответствующими числами, а затем проведем по таблице непрерывную ломаную линию, которая пройдет через все клеточки. Начиная с 1, мы сделаем сначала один шаг вправо и получим 2 в качестве второго члена последовательности; затем по диагонали налево и вниз — получим третий член $\frac{1}{2}$; следующий шаг прямо вниз даст нам четвертый член $\frac{1}{3}$; потом

движемся по диагонали вправо и вверх через $\frac{2}{2}$ к 3; вправо — к 4; по диагонали влево и вниз через $\frac{3}{2}$ и $\frac{2}{3}$ к $\frac{1}{4}$, и т. д., как показано на рис. 19. В результате движение по ломаной линии приводит к последовательности рациональных чисел

$$1,\ 2,\ \frac{1}{2},\ \frac{1}{3},\ \frac{2}{2},\ 3,\ 4,\ \frac{3}{2},\ \frac{2}{3},\ \frac{1}{4},\ \frac{1}{5},\ \frac{2}{4},\ \frac{3}{3},\ \frac{4}{2},\ 5,\ \ldots$$

Если мы выбросим теперь все дроби, у которых числитель и знаменатель имеют отличные от 1 общие делители, то останется последовательность, в которой каждое рациональное число встретится в точности один раз:

$$1,\ 2,\ \frac{1}{2},\ \frac{1}{3},\ 3,\ 4,\ \frac{3}{2},\ \frac{2}{3},\ \frac{1}{4},\ \frac{1}{5},\ 5,\ \ldots$$

Так устанавливается, что множество всех рациональных чисел является счетным. Принимая во внимание, что рациональные числа взаимно однозначно связаны с рациональными точками числовой прямой, можно также сказать, что *множество рациональных точек на числовой прямой счетно.*

Упражнения. 1) Покажите, что множество всех целых, положительных и отрицательных, чисел счетно. Покажите, что множество всех рациональных, положительных и отрицательных, чисел счетно.

2) Покажите, что если S и T — счетные множества, то множество $S + T$ (см. стр. 131) — также счетно. То же покажите для суммы трех, четырех и, вообще, n множеств; покажите, наконец, что множество, составленное посредством сложения счетного множества счетных множеств, также счетно.

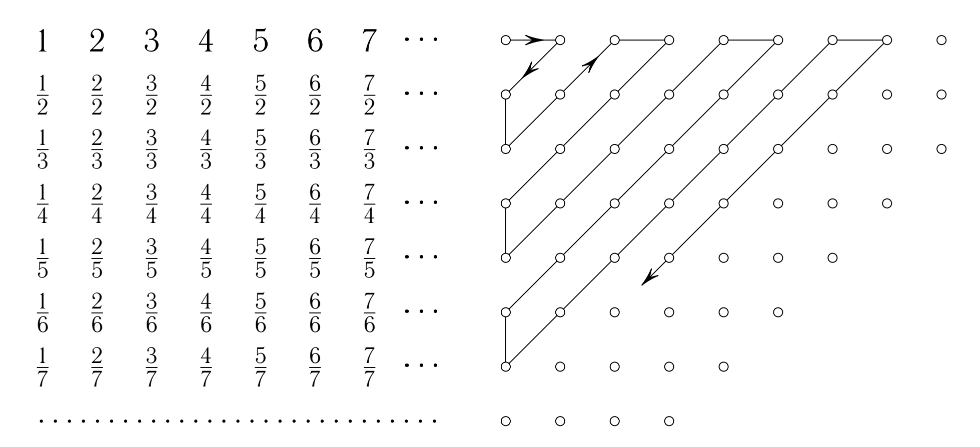


Рис. 19. Пересчет рациональных чисел

Раз оказалось, что множество рациональных чисел счетно, то могло бы возникнуть подозрение, что и *всякое* бесконечное множество также счетно, и на этом, естественно, закончился бы весь анализ бесконечного. Но это совсем не так. Тому же Кантору принадлежит открытие исключительной важности: *множество всех действительных* (рациональных и иррациональных) *чисел несчетно.* Другими словами, совокупность всех действительных чисел совершенно иного (так сказать более высокого) «типа бесконечности», чем совокупность одних только целых или одних только рациональных чисел. Принадлежащее Кантору остроумное «косвенное» доказательство этого факта стало моделью для многих иных доказательств в математике. Идея рассуждения такова. Мы исходим из допущения, что все действительные числа удалось перенумеровать, располагая их в виде последовательности, и после этого демонстрируем число, которое никак не может быть числом этой последовательности. Отсюда возникает противоречие: ведь было предположено, что все действительные числа вошли в состав последовательности, и это предположение должно быть признано ложным, если хотя бы одно число оказывается за пределами последовательности. Таким образом обнаруживается несостоятельность утверждения, что все действительные числа поддаются «пересчету», и ничего другого не остается, как только признать вместе с Кантором, что множество действительных чисел несчетно.

Однако проведем это рассуждение фактически. Допустим, что все действительные числа, представленные в виде бесконечных десятичных дробей, расположены в порядке последовательности, или списка:

$$\begin{array}{ll} \text{1-е число} & N_1{,}a_1a_2a_3a_4a_5\ldots \\ \text{2-е число} & N_2{,}b_1b_2b_3b_4b_5\ldots \\ \text{3-е число} & N_3{,}c_1c_2c_3c_4c_5\ldots \\ \ldots\ldots\ldots\ldots & \end{array}$$

где буквы N_i обозначают целую часть, а буквы a, b, c, ... представляют собой десятичные знаки, стоящие вправо от запятой. Мы допускаем, что эта последовательность дробей охватывает все действительные числа. Существенной частью доказательства является построение с помощью «диагональной процедуры» такого нового числа, относительно которого можно показать, что оно не входит в наш список.

Построим такое число. Для этого возьмем первую цифру после запятой a, какую угодно, но отличную от a_1, а также от 0 и 9 (последнее — чтобы избежать затруднений, возникающих из равенств вроде следующего: $0{,}999\cdots = 1{,}000\ldots$); затем вторую цифру b возьмем отличной от b_2, а также от 0 и 9; третью цифру c — отличной от c_3 и т. д. (Для большей определенности можно условиться в следующем: мы берем $a =$

1, если только $a_1 \neq 1$, а в случае $a_1 = 1$ возьмем $a = 2$; и аналогично для всех прочих цифр b, c, d, e, $\ldots$) Теперь рассмотрим число

$$z = 0{,}abcde\ldots$$

Это новое число z наверняка не входит в наш список; действительно, оно не равно первому числу, стоящему в списке, так как от него отличается первой цифрой после запятой, оно не равно второму числу, так как от него отличается второй цифрой после запятой, и вообще отлично от n-го числа по списку, так как от него отличается n-й цифрой после запятой. Итак, в нашем списке, составленном будто бы из всех действительных чисел, нет числа z. Значит, множество всех действительных чисел несчетно.

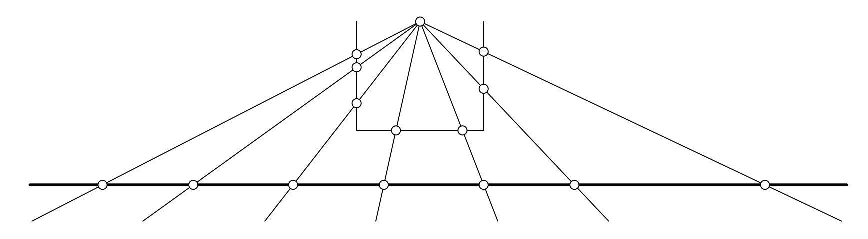


Рис. 20. Взаимно однозначное соответствие между точками согнутого интервала и точками прямой линии

Читателю может прийти в голову мысль, что несчетность континуума обусловливается неограниченной протяженностью прямой линии и что конечный отрезок прямой будет содержать лишь счетное множество точек. Чтобы убедиться в ложности такого предположения, достаточно установить, что весь числовой континуум в целом эквивалентен некоторому конечному интервалу, скажем, единичному интервалу от 0 до 1. Получить необходимое для этой цели взаимно однозначное соответствие можно, например, сгибая интервал в точках $\frac{1}{3}$ и $\frac{2}{3}$ и затем проектируя так, как показано на рис. 20. Отсюда видно, что даже конечный интервал (и, конечно, отрезок) содержит несчетное множество точек.

Упражнение. Покажите, что любой отрезок $[A, B]$ числовой прямой эквивалентен любому другому отрезку $[C, D]$ (рис. 21).

Стоит привести еще другое доказательство несчетности континуума, носящее, пожалуй, более интуитивный характер. Достаточно (принимая

во внимание последнее доказанное предложение) сосредоточить внимание на точках единичного отрезка от 0 до 1. Доказательство, впрочем, как и раньше, будет «косвенное». Предположим, что множество всех точек названного отрезка может быть расположено в виде последовательности

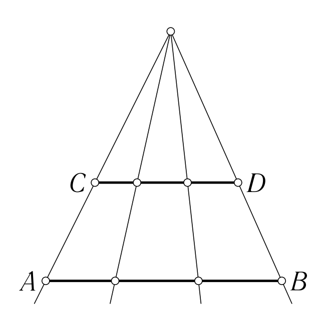


Рис. 21. Взаимно однозначное соответствие между точками двух отрезков различной длины

$$a_1,\ a_2,\ a_3,\ \ldots \qquad (1)$$

Покроем точку a_1 интервалом, длина которого пусть будет равна $\frac{1}{10}$, точку a_2 — интервалом длины $\frac{1}{10^2}$ и т. д. Если бы все точки единичного отрезка входили в последовательность (1), то весь единичный отрезок оказался бы покрытым бесконечным множеством таких отрезков (может быть, частью перекрывающихся), длины которых суть $\frac{1}{10}$, $\frac{1}{10^2}$, ... (Беды нет, если некоторые из наложенных отрезков выйдут за пределы основного единичного отрезка.) Сумма всех длин наложенных отрезков равна

$$\frac{1}{10}+\frac{1}{10^2}+\frac{1}{10^3}+\ldots=\frac{1}{10}\cdot\frac{1}{1-\frac{1}{10}}=\frac{1}{9}.$$

Итак, допущение, что последовательность (1) содержит все действительные точки единичного отрезка, приводит к заключению, что весь этот отрезок, длина которого равна 1, можно покрыть множеством промежутков с общей длиной $\frac{1}{9}$; с интуитивной точки зрения это нелепость. Это рассуждение мы позволим себе рассматривать как доказательство, хотя строго логически тут был бы нужен более глубокий анализ.

Приведенное только что рассуждение, между прочим, позволяет установить одну теорему, имеющую большое значение в современной «теории меры». Заменяя упомянутые выше промежутки меньшими промежутками — длины $\frac{\varepsilon}{10^n}$, где ε — произвольно малое положительное число, мы убедимся, что всякое счетное множество точек на прямой может быть покрыто множеством отрезков с общей длиной $\frac{\varepsilon}{9}$. Так как ε произвольно мало, то и $\frac{\varepsilon}{9}$ может быть сделано столь малым, сколь нам угодно. Пользуясь фразеологией «теории меры», мы скажем, что счетное множество точек имеет *меру нуль*.

Упражнение. Докажите аналогичную теорему для счетного множества точек на плоскости, заменяя отрезки площадями квадратов.

3. «Кардинальные числа» Кантора. Резюмируем полученные результаты. Число элементов *конечного* множества A не может равняться числу элементов другого конечного множества B, если A содержит

больше элементов, чем B. Но если мы заменим понятие «множеств, имеющих одно и то же конечное число элементов» более общим понятием «эквивалентных множеств», то — в случае бесконечных множеств — предыдущее утверждение уже не будет справедливо: множество всех целых чисел содержит «больше» элементов, чем множество всех четных чисел, а множество всех рациональных чисел — «больше» элементов, чем множество всех целых чисел; и, однако, как мы видели, все эти множества эквивалентны. Можно было бы заподозрить, что *все* бесконечные множества между собой эквивалентны, но Кантор опроверг это предположение: существует множество — континуум действительных чисел, — которое не эквивалентно никакому счетному множеству.

Итак, существует по меньшей мере два различных «типа бесконечности»: счетная бесконечность натуральных чисел и несчетная бесконечность континуума. Если два множества A и B, конечные или бесконечные, эквивалентны, мы скажем, что им соответствует одно и то же *кардинальное число* (или *мощность*). В случае конечных множеств кардинальное число сводится к обыкновенному натуральному числу, но понятие кардинального числа носит более общий характер. Далее, если случится, что множество A эквивалентно некоторому подмножеству (части) множества B, но само B неэквивалентно ни множеству A, ни какой бы то ни было его части, то говорят, следуя Кантору, что множеству B соответствует *большее* кардинальное число, чем множеству A. Это употребление термина «число» также согласуется с обычным употреблением в случае конечных множеств. Множество целых чисел есть подмножество множества всех действительных чисел, тогда как множество действительных чисел не эквивалентно ни множеству целых чисел, ни какому бы то ни было его подмножеству (оно ни счетное, ни конечное). Значит, по данному определению, континууму действительных чисел соответствует большее кардинальное число, чем множеству натуральных чисел.

* Кантор показал фактически, как можно построить бесконечную последовательность бесконечных множеств, которым соответствуют все бо́льшие и бо́льшие кардинальные числа. Так как можно исходить из множества натуральных чисел, то достаточно показать, что, *каково бы ни было данное множество A, можно построить другое множество B, у которого кардинальное число будет больше, чем у A.* Вследствие большой общности этой теоремы доказательство ее по неизбежности несколько абстрактно. Множество B мы определяем как множество, элементами которого являются все возможные подмножества множества A. Говоря о «подмножествах» A, мы в данном случае имеем в виду не только «правильные подмножества» A, но не исключаем и самого множества A, а также «пустого» множества $\varnothing$, не содержащего никаких элементов. (Так, если A состоит из трех целых чисел 1, 2, 3, то B содержит 8 различных элементов $\{1, 2, 3\}$, $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$, $\{1\}$, $\{2\}$, $\{3\}$ и $\varnothing$.) Каждый

элемент множества B сам есть множество, состоящее из каких-то элементов множества A. Допустим теперь, что B эквивалентно A или некоторому подмножеству A, т. е. что существует некоторое правило, приводящее во взаимно однозначное соответствие элементы A или некоторого подмножества A со всеми элементами B, т. е. всеми подмножествами A:

$$a \leftrightarrow S_a, \tag{2}$$

где через S_a обозначено то подмножество A, которому соответствует элемент a множества A. Мы придем к противоречию, если укажем некоторый элемент B, т. е. некоторое подмножество T множества A, которому не может соответствовать никакой элемент a. Чтобы построить подмножество T, заметим прежде всего, что для всякого элемента x из A существуют две возможности: либо множество S_x, сопоставляемое зависимостью (2) элементу x, содержит элемент x, либо не содержит. Мы определим T как *подмножество A, состоящее из всех таких элементов x, что S_x не содержит x.* Определенное таким образом множество T отличается от всякого S_a по крайней мере элементом a, так как если S_a содержит a, то T не содержит a, а если S_a не содержит a, то T содержит a. Итак, T не включено в соответствие (2). Это и показывает, что невозможно установить взаимно однозначное соответствие между элементами A (или некоторого подмножества A) и элементами B. Но соотношение

$$a \leftrightarrow \{a\}$$

устанавливает взаимно однозначное соответствие между всеми элементами A и подмножеством B, состоящим из одноэлементных подмножеств A. Значит, по данному выше определению, множеству B соответствует большее кардинальное число, чем множеству A.

* **Упражнение.** Если множество A содержит n элементов, то определенное выше множество B содержит 2^n элементов. Если A есть множество натуральных чисел, то B эквивалентно континууму действительных чисел, заключенных между 0 и 1. (*Указание*: сопоставьте каждому подмножеству A символ, состоящий из последовательности — конечной в первом примере, бесконечной во втором —

$$a_1 a_2 a_3 \ldots,$$

где $a_n = 1$ или 0, смотря по тому, принадлежит или не принадлежит n-й элемент A рассматриваемому подмножеству.)

Могло бы показаться легкой задачей построить множество точек, обладающее бо́льшим кардинальным числом, чем множество точек единичного отрезка. Казалось бы, что квадрат со стороной 1, как «двумерная» фигура, должен содержать «больше» точек, чем «одномерный» отрезок. Но, как это ни странно, дело обстоит иначе: *кардинальное число точек квадрата в точности равно кардинальному числу точек отрезка.* Для доказательства достаточно установить взаимно однозначное соответствие между точками квадрата и точками отрезка. Постараемся это сделать.

Если (x, y) есть какая-нибудь точка единичного квадрата, то ее координаты x и y могут быть представлены в виде десятичных разложений

$$x = 0{,}a_1 a_2 a_3 a_4 \ldots, \qquad y = 0{,}b_1 b_2 b_3 b_4 \ldots,$$

причем пусть будет условлено (ради избежания всяких сомнений), что, например, число $\frac{1}{4}$ будет записываться в виде 0,25000..., а не в виде 0,24999...
Названной точке квадрата (x, y) мы сопоставим точку единичного отрезка

$$z = 0{,}a_1b_1a_2b_2a_3b_3a_4b_4\ldots.$$

Очевидно, различным точкам квадрата (x, y) и (x', y') сопоставляются различные же точки отрезка z и z'; это и значит, что кардинальное число множества точек квадрата не превышает кардинального числа множества точек отрезка.

(Собственно говоря, в данном случае построено взаимно однозначное соответствие между множеством всех точек квадрата и некоторым подмножеством точек отрезка: никакая точка квадрата не будет соответствовать, например, точке отрезка 0,2140909090..., так как мы условились писать 0,25000..., а не 0,24999... Но можно слегка видоизменить построение таким образом, чтобы действительно осуществлялось взаимно однозначное соответствие между множеством всех точек квадрата и множеством всех точек отрезка.)

Аналогичнее рассуждение показывает, что кардинальнее число точек куба не превышает кардинального числа точек отрезка.

Все эти результаты, казалось бы, стоят в противоречии с интуитивным представлением о «размерности». Но нужно обратить внимание на то, что вводимые нами соответствия не являются «непрерывными»; когда мы перемещаемся по отрезку от 0 к 1 непрерывно, соответствующие точки в квадрате не образуют непрерывной кривой, а будут появляться в порядке совершенно «хаотическом». Размерность множества точек зависит не только от кардинального числа точек, но и от того, как они расположены в пространстве. Мы вернемся к этому вопросу в главе V.

4. Косвенный метод доказательства. Теория кардинальных чисел представляет собой лишь один из аспектов общей теории множеств, созданной Кантором несмотря на суровую критику со стороны наиболее выдающихся математиков того времени. Многие из критиков, например Пуанкаре и Кронекер, возражали против неопределенности общего понятия «множества» и против неконструктивного характера рассуждений, применявшихся при определении некоторых множеств.

Возражения против неконструктивных рассуждений относятся к тем доказательствам, которые можно было бы назвать «существенно косвенными». Сами по себе «косвенные» доказательства есть самый обыкновенный элемент математического мышления: желая установить истинность предложения A, мы вначале допускаем, что справедливо иное предложение A', противоположное A; затем некоторая цепь рассуждений приводит нас к утверждению, противоречащему A', и тем самым обнаруживается несостоятельность предложения A'. Тогда на базе основного логического принципа «исключенного третьего» из ложности A' следует истинность A.

В разных местах этой книги читатель найдет ряд таких примеров,

для которых косвенное доказательство легко может быть превращено в прямое, но «косвенная» форма создает преимущества краткости и освобождает от рассмотрения подробностей, имеющих второстепенный интерес с точки зрения поставленной ближайшей цели. Но попадаются и такие теоремы, для которых до настоящего времени не удалось дать иных доказательств, кроме косвенных. О некоторых из этих теорем можно даже сказать, что, по-видимому, по самой их природе прямые, конструктивные их доказательства принципиально невозможны. Сюда относится, например, теорема, приведенная на стр. 102. Не раз бывали случаи в истории математики, когда все усилия математиков были направлены в сторону *построения* («конструкции») решения тех или иных проблем, разрешимость которых предполагалось установить, а затем кто-нибудь приходил, если так можно выразиться, со стороны и ликвидировал все трудности с помощью «косвенного» неконструктивного рассуждения.

Когда речь идет о доказательстве существования объекта определенного типа, то имеется существенное различие между тем, чтобы построить осязаемый пример объекта, и тем, чтобы доказать, что из несуществования объекта можно вывести противоречивые заключения. В первом случае получается осязаемый объект, во втором — ничего, кроме противоречия. Не так давно некоторые математики (весьма заслуженные) провозгласили более или менее полное устранение из математики всех неконструктивных доказательств. Даже если бы выполнение этой программы признать желательным, необходимо указать, что это повлекло бы за собой в настоящую эпоху чрезвычайные усложнения, и можно было бы даже опасаться, что в процессе совершающихся потрясений подверглись бы разрушению существенные части организма математики. Поэтому нечего удивляться, что школа «интуиционистов», принявшая упомянутую программу, встретила упорное сопротивление, и что даже наиболее ортодоксальные интуиционисты не всегда в состоянии жить согласно своим убеждениям.[1]

5. Парадоксы бесконечного. Хотя бескомпромиссная позиция, занятая интуиционистами, с точки зрения большинства математиков является слишком экстремистской, волей-неволей приходится согласиться, что для внешне прекрасной теории бесконечных множеств возникла серьезная угроза, когда в пределах самой этой теории обнаружились совершенно явные логические парадоксы. Очень скоро было замечено, что неограниченная свобода в пользовании понятием «множество» неиз-

[1] Об интуиционизме и выросшем из него конструктивном направлении в математике и логике, на исчерпывающую характеристику которых никак не претендуют эти строки, см., например, [11] и [15] в списке литературы в конце книги (номера по которому всюду указываются в квадратных скобках). — *Прим. ред.*

бежно ведет к противоречиям. Мы приведем здесь один из парадоксов, обнаруженный Бертраном Расселом. Вот в чем он заключается.

Как правило, множества не содержат себя в качестве элемента. Например, множество A всех целых чисел содержит в качестве элементов только целые числа; так как само A не есть целое число, а есть *множество* целых чисел, то A себя в качестве элемента не содержит. Условимся называть такие множества «ординарными». Но могут существовать и такие множества, которые содержат себя в качестве элемента. Рассмотрим, например, множество S, определенное следующим образом: «S содержит в качестве элементов все множества, которые можно определить посредством предложения, содержащего меньше двадцати слов». Так как само множество S определяется предложением, содержащим меньше двадцати слов, то выходит, что оно является элементом множества S. Такие множества назовем «экстраординарными». Как бы то ни было, большинство множеств — ординарные; попробуем не иметь дела с дурно ведущими себя экстраординарными множествами и будем рассматривать только *множество всех ординарных множеств*. Обозначим его буквой C. Каждый элемент C есть множество, притом ординарное множество. Но вот возникает вопрос: *а само множество C — ординарное или экстраординарное*? Несомненно, оно должно быть или тем, или другим.[1] Если C — ординарное множество, то оно содержит себя в качестве элемента, так как C определено как множество *всех* ординарных множеств. Раз дело обстоит так, значит, C — экстраординарное множество, так как экстраординарными, согласно определению, названы множества, содержащие себя в качестве элемента. Получается противоречие. Значит, C должно быть экстраординарным множеством. Но тогда множество C содержит в качестве элемента себя, т. е. оно есть экстраординарное множество, а это противоречит определению C как множества всех ординарных множеств. Итак, мы видим, что уже одно только допущение существования множества C внутренне противоречиво.

6. Основания математики. Парадоксы вроде вышеприведенного побудили Рассела и других подвергнуть систематическому изучению основания математики и логики. Конечная цель этих исследований заключается в создании для математических рассуждений такой логической базы, относительно которой можно было бы доказать, что она свободна от возможных противоречий, и которая вместе с тем была бы достаточно обширной, чтобы из нее можно было путем дедукции

[1] Получаемое далее противоречие может быть выведено и без использования так называемого закона исключенного третьего, подразумеваемого в этой фразе. См., например, Френкель А. и Бар-Хиллел И. Основания теории множеств. — М.: Мир, 1966, гл. I, § 2. — *Прим. ред.*

вывести все, что в математике признается существенным, или хотя бы многое из того. Поскольку такой самонадеянной цели достигнуть не удавалось (а может быть, ее и нельзя достигнуть), математическая логика как особый предмет привлекала внимание все возрастающего числа исследователей. Многие относящиеся сюда проблемы необходимо признать крайне трудными, хотя формулировки их вполне просты. В качестве примера назовем *гипотезу континуума*, утверждающую, что не существует множества, для которого кардинальное число больше, чем кардинальное число множества натуральных чисел, но меньше, чем кардинальное число множества действительных чисел. Из этой гипотезы можно вывести много интересных следствий, но сама гипотеза до наших дней не была ни доказана, ни опровергнута. Впрочем, не так давно[1] Курт Гёдель доказал, что если система обычных постулатов, лежащих в основе теории множеств, не содержит противоречий, то в таком случае расширенная система постулатов, получающаяся при добавлении континуум-гипотезы, также не содержит противоречий. Вопросы, рассматриваемые в математической логике, в конечном счете упираются в один основной вопрос: что понимать под существованием в математике? К счастью, существование самой математики не зависит от того, найден ли удовлетворительный ответ на этот вопрос. Школа «формалистов», во главе которой стоял великий математик Гильберт, утверждает, что в математике «существование» означает «свободу от противоречия». Если принять эту точку зрения, то очередной и необходимой задачей является как раз построение системы постулатов, из которых всю математику можно было бы вывести путем логической дедукции, и доказательство того, что эти постулаты не могут привести ни к какому противоречию. Недавние результаты Гёделя и других как будто бы показывают, что такая программа, по крайней мере в той форме, в какой она была намечена самим Гильбертом, не может быть осуществлена. Весьма многозначительно то обстоятельство, что гильбертова теория формализированного построения математики существенно опирается на интуитивные процедуры. Тем или иным путем, в открытой или в скрытой форме, даже прикрытая самым безупречным формалистическим, логическим, аксиоматическим одеянием, конструктивная интуиция всегда остается самым жизненным элементом в математике.[2]

---

[1] 1940 г. А в 1963 г. американским математиком П. Коэном доказана независимость континуум-гипотезы от принятой Гёделем системы аксиом теории множеств. — *Прим. ред.*

[2] Подробнее об этих вопросах см. [11] и [38]. — *Прим. ред.*

§ 5. Комплексные числа

1. Возникновение комплексных чисел. По ряду причин возникла потребность в расширении понятия числа даже за пределы континуума действительных чисел — посредством введения так называемых комплексных чисел. Необходимо ясно представлять себе, что все подобного рода расширения и нововведения приходят отнюдь не в результате чьих-то индивидуальных усилий. Скорее их можно рассматривать как итог некоторой постепенной и исполненной колебаний эволюции, в которой не следует преувеличивать роль отдельных личностей. Одной из причин, которые обусловили появление и употребление отрицательных и дробных чисел, было стремление к большей свободе в формальных вычислениях. Только к концу средневековья математики стали терять ощущение беспокойства и неуверенности, с которым они оперировали этими понятиями, тогда как ничего подобного не наблюдалось в отношении таких интуитивно ясных и конкретно воспринимаемых понятий, как понятие натурального числа.

Простейшая процедура, требующая применения комплексных чисел, есть *решение квадратных уравнений*. Напомним, как обстояло дело с линейным уравнением $ax = b$, когда нужно было определить удовлетворяющее ему значение неизвестной величины x. Решение имеет вид $x = \dfrac{b}{a}$, и введение дробных чисел как раз обусловливается требованием, чтобы всякое линейное уравнение с целыми коэффициентами (при $a \neq 0$) было разрешимо. Уравнения вроде

$$x^2 = 2 \tag{1}$$

не имеют решения в области рациональных чисел, но имеют таковое в расширенном поле всех действительных чисел. Но даже поле действительных чисел недостаточно обширно, чтобы в нем можно было построить полную и законченную теорию квадратных уравнений. Например, следующее очень простое уравнение

$$x^2 = -1 \tag{2}$$

не имеет действительных решений, так как квадрат действительного числа никак не может быть отрицательным. Нам приходится или удовольствоваться тем положением, что такие простые уравнения неразрешимы, или следовать по уже знакомому пути — расширять числовую область и вводить новые числа, с помощью которых удастся решить уравнение. Именно это самое и делается, когда вводят новый символ i и принимают, *в качестве определения*, что $i^2 = -1$. Разумеется, этот объект — «мнимая единица» — не имеет ничего общего с числом как орудием *счета*. Это — отвлеченный *символ*, подчиненный основному закону $i^2 = -1$, и ценность его зависит исключительно от того, будет ли достигнуто

в результате его введения действительно полезное расширение числовой системы.

Так как мы хотим складывать и умножать с помощью символа i так же, как с обыкновенными числами, то естественно пользоваться символами вроде $2i$, $3i$, $-i$, $2+5i$, вообще, $a+bi$, где a и b — действительные числа. Раз эти символы должны подчиняться коммутативному, ассоциативному и дистрибутивному законам, то должны быть возможны, например, такие вычисления:

$$(2+3i)+(1+4i)=(2+1)+(3+4)i=3+7i;$$
$$(2+3i)\cdot(1+4i)=2+8i+3i+12i^2=(2-12)+(8+3)i=-10+11i.$$

Руководствуясь этими соображениями, мы начинаем систематическое изложение теории комплексных чисел со следующего *определения*: символ вида $a+bi$, где a и b — два действительных числа, носит название *комплексного числа с действительной частью a и мнимой частью b*. Операции сложения и умножения совершаются над этими числами так, как будто бы i было обыкновенное действительное число, однако с условием заменять i^2 на -1. Точнее говоря, сложение и умножение определяются по формулам

$$\left.\begin{aligned}(a+bi)+(c+di)&=(a+c)+(b+d)i,\\ (a+bi)(c+di)&=(ac-bd)+(ad+bc)i.\end{aligned}\right\} \qquad (3)$$

В частности, мы получаем

$$(a+bi)(a-bi)=a^2-abi+abi-b^2i^2=a^2+b^2. \qquad (4)$$

Основываясь на этих определениях, легко проверить, что для комплексных чисел справедливы коммутативный, ассоциативный и дистрибутивный законы. Далее, не только сложение и умножение, но также и вычитание и деление, будучи применены к двум комплексным числам, приводят снова к комплексным числам того же вида $a+bi$, так что комплексные числа образуют поле (см. стр. 75):

$$\left.\begin{aligned}(a+bi)-(c+di)&=(a-c)+(b-d)i,\\ \frac{a+bi}{c+di}=\frac{(a+bi)(c-di)}{(c+di)(c-di)}&=\left(\frac{ac+bd}{c^2+d^2}\right)+\left(\frac{bc-ad}{c^2+d^2}\right)i.\end{aligned}\right\} \qquad (5)$$

(Второе равенство теряет смысл, если $c+di=0+0i$, так как тогда $c^2+d^2=0$. Значит, и на этот раз *нужно исключить деление на нуль*, т. е. на $0+0i$.) Например,

$$(2+3i)-(1+4i)=1-i,$$
$$\frac{2+3i}{1+4i}=\frac{2+3i}{1+4i}\cdot\frac{1-4i}{1-4i}=\frac{2-8i+3i+12}{1+16}=\frac{14}{17}-\frac{5}{17}i.$$

Поле комплексных чисел включает поле действительных чисел в качестве «подполя», так как комплексное число $a+0i$ отождествляется с

действительным числом a. Заметим, с другой стороны, что комплексное число вида $0 + bi = bi$ называется «чисто мнимым».

Упражнения. 1) Представьте $\dfrac{(1+i)(2+i)(3+i)}{(1-i)}$ в форме $a + bi$.

2) Представьте

$$\left(-\frac{1}{2} + i\,\frac{\sqrt{3}}{2}\right)^3$$

в форме $a + bi$.

3) Представьте в форме $a + bi$ следующие выражения:

$$\frac{1+i}{1-i}, \quad \frac{1+i}{2-i}, \quad \frac{1}{i^5}, \quad \frac{1}{(-2+i)(1-3i)}, \quad \frac{(4-5i)^2}{(2-3i)^2}.$$

4) Вычислите $\sqrt{5+12i}$. (*Указание*: напишите $\sqrt{5+12i} = x + yi$, возведите в квадрат и приравняйте действительные части и мнимые части.)

Вводя символ i, мы расширили поле действительных чисел и получили поле символов $a + bi$, в котором квадратное уравнение

$$x^2 = -1$$

имеет два решения: $x = i$ и $x = -i$. В самом деле, согласно определению, $i \cdot i = (-i)(-i) = i^2 = -1$. Нужно сказать, что мы приобрели гораздо больше: можно легко проверить, что теперь *каждое* квадратное уравнение

$$ax^2 + bx + c = 0 \tag{6}$$

становится разрешимым. В самом деле, выполняя над равенством (6) ряд преобразований, мы получаем:

$$x^2 + \frac{b}{a}x = -\frac{c}{a},$$

$$x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} = \frac{b^2}{4a^2} - \frac{c}{a},$$

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2},$$

$$x + \frac{b}{2a} = \frac{\pm\sqrt{b^2 - 4ac}}{2a},$$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{7}$$

Заметим теперь, что если $b^2 - 4ac \geqslant 0$, то $\sqrt{b^2 - 4ac}$ есть обыкновенное действительное число и корни уравнения (6) действительные; если же $b^2 - 4ac < 0$, то тогда $4ac - b^2 > 0$, и следовательно, $\sqrt{b^2 - 4ac} = \sqrt{-(4ac - b^2)} = \sqrt{4ac - b^2} \cdot i$, так что уравнение (6) имеет в качестве корней мнимые числа. Так, например, уравнение

$$x^2 - 5x - 6 = 0$$

имеет действительные корни $x = \dfrac{5 \pm \sqrt{25 - 24}}{2} = \dfrac{5 \pm 1}{2} = 3$ или 2, тогда как уравнение

$$x^2 - 2x + 2 = 0$$

имеет мнимые корни $x = \frac{2 \pm \sqrt{4-8}}{2} = \frac{2 \pm 2i}{2} = 2 = 1 + i$ или $1 - i$.

2. Геометрическое представление комплексных чисел. Уже в XVI столетии в математических работах появляются квадратные корни из отрицательных чисел в формулах, дающих решения квадратных уравнений. Но в те времена математики затруднились бы объяснить точный смысл этих выражений, к которым относились почти с суеверным трепетом. Сам термин «мнимый» до сих пор напоминает нам о том, что эти выражения рассматривались как нечто искусственное, лишенное реального значения. И только в начале XIX в., когда уже выяснилась роль комплексных чисел в различных областях математики, было дано очень простое геометрическое истолкование комплексных чисел и операций с ними, и этим был положен конец сомнениям в возможности их законного употребления. Конечно, с современной точки зрения, формальные операции с комплексными числами полностью оправдываются на основе формальных определений, так что геометрическое представление логически не является необходимым. Но такое представление, предложенное почти одновременно Весселем (1745–1818), Арганом (1768–1822) и Гауссом, позволило рассматривать комплексные числа и действия с ними как нечто вполне естественное с интуитивной точки зрения и, кроме того, имеющее чрезвычайно большое значение в приложениях комплексных чисел как в самой математике, так и в математической физике.

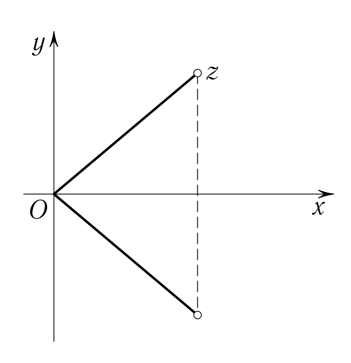


Рис. 22. Геометрическое представление комплексных чисел. Точка z имеет прямоугольные координаты x, y

Геометрическая интерпретация комплексных чисел заключается в том, что комплексному числу $z = x + yi$ сопоставляется точка на плоскости с координатами x, y. Именно, действительная часть числа мыслится как x-координата, а мнимая — как y-координата. Таким образом устанавливается взаимно однозначное соответствие между комплексными числами и точками «числовой плоскости», подобно тому как нами было установлено раньше (см. § 2) соответствие между действительными числами и точками «числовой оси». Точкам на оси x в числовой плоскости соответствуют действительные числа $z = x + 0i$, тогда как точкам на оси y — чисто мнимые числа $z = 0 + yi$.

Если

$$z = x + yi$$

есть какое-то комплексное число, то мы называем число

$$\overline{z} = x - yi$$

сопряженным с числом z. В числовой плоскости точка $\overline{z}$ получается из точки z посредством зеркального отражения относительно оси x. Если мы условимся расстояние точки z от начала обозначать через ρ, то на основании теоремы Пифагора

$$\rho^2 = x^2 + y^2 = (x + yi)(x - yi) = z \cdot \overline{z}.$$

Действительное число $\rho = \sqrt{x^2 + y^2}$ называется *модулем* z и обозначается

$$\rho = |z|.$$

Если z лежит на действительной оси, то модуль совпадает с абсолютной величиной z. Комплексные числа с модулем 1 изображаются точками, лежащими на «единичной окружности» с центром в начале и радиусом 1.

Если $|z| = 0$, *то* $z = 0$. Это следует из определения $|z|$ как расстояния точки z от начала. Далее, *модуль произведения двух комплексных чисел равен произведению модулей*:

$$|z_1 \cdot z_2| = |z_1| \cdot |z_2|.$$

Это вытекает как следствие из более общей теоремы, которая будет доказана на стр. 117.

Упражнения. 1) Докажите последнюю теорему, исходя непосредственно из определения умножения двух комплексных чисел $z_1 = x_1 + y_1 i$ и $z_2 = x_2 + y_2 i$.

2) Пользуясь тем обстоятельством, что произведение двух *действительных* чисел равно нулю в том и только том случае, если один из множителей *равен нулю*, докажите соответствующую теорему для *комплексных чисел*. (*Указание*: основывайтесь при доказательстве на двух последних теоремах.)

Согласно определению сложения двух комплексных чисел $z_1 = x_1 + y_1 i$ и $z_2 = x_2 + y_2 i$, мы имеем

$$z_1 + z_2 = (x_1 + x_2) + (y_1 + y_2)i.$$

Таким образом, точка $z_1 + z_2$ изображается в числовой плоскости четвертой вершиной параллелограмма, у которого тремя первыми вершинами являются точки 0, z_1, z_2. Это простой способ построения суммы двух комплексных чисел ведет ко многим важным следствиям. Из него мы заключаем, что *модуль суммы двух комплексных чисел не превышает суммы модулей* (ср. стр. 76):

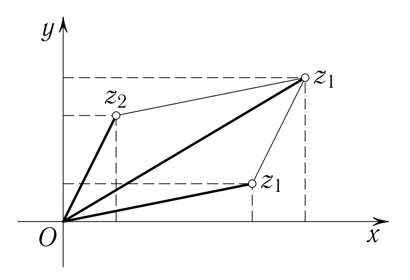


Рис. 23. Сложение комплексных чисел по правилу параллелограмма

$$|z_1 + z_2| \leqslant |z_1| + |z_2|.$$

Достаточно сослаться на то, что длина стороны треугольника не превышает суммы длин двух других сторон.

Упражнение. В каких случаях имеет место равенство $|z_1 + z_2| = |z_1| + |z_2|$?

Угол между положительным направлением оси x и отрезком Oz называется *аргументом* z и обозначается буквой φ (см. рис. 22). Числа z и $\overline{z}$ имеют один и тот же модуль

$$|\overline{z}| = |z|,$$

но их аргументы противоположны по знаку:

$$\overline{\varphi} = -\varphi.$$

Конечно, аргумент z определяется не однозначно, так как к нему можно прибавлять или из него вычитать любой угол, кратный 360°, не изменяя направления отрезка Oz. Итак, углы

$$\begin{array}{llll} \varphi, \quad \varphi + 360^\circ, & \varphi + 720^\circ, & \varphi + 1080^\circ, & \ldots \\ \quad\;\; \varphi - 360^\circ, & \varphi - 720^\circ, & \varphi - 1080^\circ, & \ldots \end{array}$$

графически дают один и тот же аргумент. Так как, согласно определению синуса и косинуса,

$$x = \rho\cos\varphi, \qquad y = \rho\sin\varphi,$$

то любое комплексное число z выражается через его модуль и аргумент следующим образом:

$$z = x + yi = \rho(\cos\varphi + i\sin\varphi). \tag{8}$$

Например,

в случае		$z = i$	мы имеем		$\rho = 1$,	$\varphi = 90^\circ$,
»	»	$z = 1 + i$	»	»	$\rho = \sqrt{2}$,	$\varphi = 45^\circ$,
»	»	$z = 1 - i$	»	»	$\rho = \sqrt{2}$,	$\varphi = -45^\circ$,
»	»	$z = -1 + \sqrt{3}i$	»	»	$\rho = 2$,	$\varphi = 120^\circ$,

так что

$$\begin{aligned} i &= 1\,(\cos 90^\circ + i\sin 90^\circ), \\ 1 + i &= \sqrt{2}\,(\cos 45^\circ + i\sin 45^\circ), \\ 1 - i &= \sqrt{2}\,(\cos(-45^\circ) + i\sin(-45^\circ)), \\ -1 + \sqrt{3}i &= 2\,(\cos 120^\circ + i\sin 120^\circ). \end{aligned}$$

Читатель может проверить эти утверждения посредством подстановки числовых значений тригонометрических функций.

Тригонометрическим представлением (8) очень полезно воспользоваться, чтобы уяснить себе геометрический смысл умножения двух комплексных чисел. Если

$$z = \rho\,(\cos\varphi + i\sin\varphi)$$

и
$$z' = \rho'(\cos\varphi' + i\sin\varphi'),$$
то
$$zz' = \rho\rho'\big\{(\cos\varphi\cos\varphi' - \sin\varphi\sin\varphi') + (\cos\varphi\sin\varphi' + \sin\varphi\cos\varphi')i\big\}.$$
Но, в силу основных теорем сложения синуса и косинуса,
$$\cos\varphi\cos\varphi' - \sin\varphi\sin\varphi' = \cos(\varphi + \varphi'),$$
$$\cos\varphi\sin\varphi' + \sin\varphi\cos\varphi' = \sin(\varphi + \varphi').$$
Итак,
$$zz' = \rho\rho'\{\cos(\varphi + \varphi') + i\sin(\varphi + \varphi')\}. \tag{9}$$
В правой части последнего равенства мы видим написанное в тригонометрической форме комплексное число с модулем $\rho\rho'$ и аргументом $\varphi + \varphi'$. Значит, мы можем отсюда заключить, что *при умножении двух комплексных чисел их модули перемножаются, а аргументы складываются* (рис. 24). Таким образом, мы видим, что умножение комплексных чисел как-то связано с *вращением.*

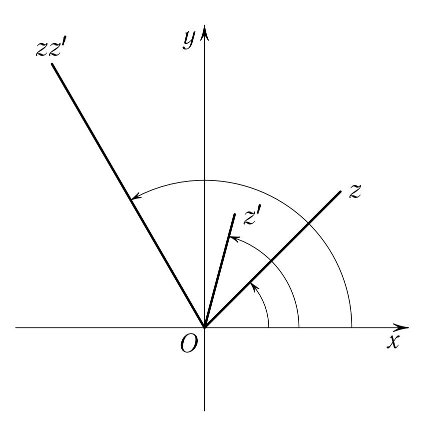


Рис. 24. Умножение комплексных чисел: аргументы складываются, модули перемножаются

Установим точнее, в чем тут дело. Назовем направленный отрезок, идущий из начала в точку z, *вектором* точки z; тогда модуль $\rho = |z|$ есть его длина. Пусть z' — какая-нибудь точка единичной окружности, так что $\rho' = 1$. В таком случае умножение z на z' просто поворачивает вектор z на угол φ'. Если же $\rho' \neq 1$, то, помимо вращения, длина вектора должна быть умножена на ρ'. Рекомендуем читателю самостоятельно проиллюстрировать эти факты, умножая различные комплексные числа на $z_1 = i$ (вращение на $90°$); $z_2 = -i$ (тоже вращение на $90°$, но в обратном направлении); $z_3 = 1 + i$ и $z_4 = 1 - i$.

Формула (9) в особенности представляет интерес, если $z = z'$; в этом случае имеем:
$$z^2 = \rho^2(\cos 2\varphi + i\sin 2\varphi).$$
Умножая снова на z, будем иметь
$$z^3 = \rho^3(\cos 3\varphi + i\sin 3\varphi);$$
и, вообще, для любого n, повторяя операцию, получим
$$z^n = \rho^n(\cos n\varphi + i\sin n\varphi). \tag{10}$$

В частности, если точка z находится на *единичной окружности*, так что $\rho = 1$, мы приходим к формуле, открытой французским математиком А. де Муавром (1667–1754):

$$(\cos\varphi + i\sin\varphi)^n = \cos n\varphi + i\sin n\varphi. \tag{11}$$

Эта формула — одно из самых замечательных и полезных соотношений в элементарной математике. Поясним это примером. Возьмем $n = 3$ и разложим левую часть по формуле бинома

$$(u + v)^3 = u^3 + 3u^2v + 3uv^2 + v^3.$$

Тогда получим:

$$\cos 3\varphi + i\sin 3\varphi = \cos^3 3\varphi - 3\cos\varphi\sin^2\varphi + i(3\cos^2\varphi\sin\varphi - \sin^3\varphi).$$

Одно такое *комплексное равенство равносильно двум равенствам, связывающим действительные числа.* В самом деле, если два комплексных числа равны, то в отдельности равны их действительные части и их мнимые части. Итак, можно написать

$$\cos 3\varphi = \cos^3\varphi - 3\cos\varphi\sin^2\varphi, \quad \sin 3\varphi = 3\cos^2\varphi\sin\varphi - \sin^3\varphi.$$

Пользуясь затем соотношением

$$\cos^2\varphi + \sin^2\varphi = 1,$$

получим окончательно:

$$\cos 3\varphi = \cos^3\varphi - 3\cos\varphi(1 - \cos^2\varphi) = 4\cos^3\varphi - 3\cos\varphi,$$
$$\sin 3\varphi = -4\sin^3\varphi + 3\sin\varphi.$$

Подобного рода формулы, выражающие $\sin n\varphi$ и $\cos n\varphi$ соответственно через $\sin\varphi$ и $\cos\varphi$, легко получить при каком угодно целом значении n.

Упражнения. 1) Напишите аналогичные формулы для $\sin 4\varphi$ и $\cos 4\varphi$.

2) Предполагая, что точка z находится на единичном круге: $z = \cos\varphi + i\sin\varphi$, покажите, что $\dfrac{1}{z} = \cos\varphi - i\sin\varphi$.

3) Без вычислений установите, что модуль числа $\dfrac{a+bi}{a-bi}$ равен единице.

4) Докажите: если z_1 и z_2 — два комплексных числа, то аргумент $z_1 - z_2$ равен углу между положительным направлением действительной оси и вектором, идущим от z_2 к z_1.

5) Дан треугольник с вершинами z_1, z_2, z_3; установите геометрический смысл аргумента числа $\dfrac{z_1 - z_2}{z_1 - z_3}$.

6) Докажите, что отношение двух комплексных чисел с одинаковым аргументом есть действительное число.

7) Докажите, что если аргументы чисел $\dfrac{z_3 - z_1}{z_3 - z_2}$ и $\dfrac{z_4 - z_1}{z_4 - z_2}$ равны между собой, то четыре точки z_1, z_2, z_3, z_4 лежат на окружности или на прямой линии, и обратно.

8) Докажите: четыре точки z_1, z_2, z_3, z_4 лежат на окружности или на прямой линии, если число

$$\frac{z_3 - z_1}{z_3 - z_2} : \frac{z_4 - z_1}{z_4 - z_2}$$

действительное.

3. Формула Муавра и корни из единицы. Под корнем n-й степени из числа a мы понимаем всякое такое число b, что $b^n = a$. В частности, число 1 имеет два квадратных корня: 1 и -1, так как $1^2 = (-1)^2 = 1$. Число 1 имеет один *действительный* кубический корень, именно 1, тогда как оно же имеет четыре корня четвертой степени: два действительных, 1 и -1, и два мнимых: i и $-i$. Эти факты наводят на мысль, что в комплексной области должно существовать еще два кубических корня из 1 (а всего кубических корней тогда будет три). С помощью формулы Муавра мы покажем, что эта догадка справедлива.

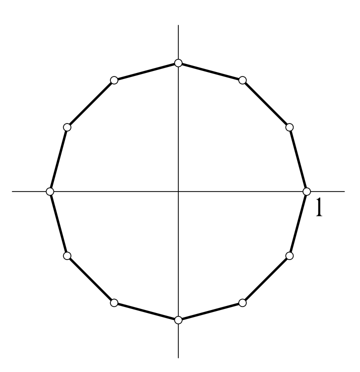


Рис. 25. Двенадцать корней двенадцатой степени из единицы

Мы убедимся, что в *поле комплексных чисел существует ровно n корней степени n из 1. Эти корни изображаются вершинами правильного n-угольника, вписанного в единичный круг и имеющего точку 1 в качестве одной из вершин.*

Сказанное почти ясно из рис. 25 (соответствующего случаю $n = 12$). Первая вершина многоугольника есть 1. Следующая есть

$$\alpha = \cos\frac{360^\circ}{n} + i\sin\frac{360^\circ}{n}, \tag{12}$$

так как аргумент должен равняться n-й части угла в 360°. Еще следующая вершина есть $\alpha \cdot \alpha = \alpha^2$, так как мы получим ее, вращая вектор α на угол $\dfrac{360^\circ}{n}$. Дальше получаем вершину α^3 и т. д.; после n шагов возвращаемся снова к вершине 1, т. е. получаем

$$\alpha^n = 1,$$

что следует также из формулы (11), так как

$$\left(\cos\frac{360^\circ}{n} + i\sin\frac{360^\circ}{n}\right) = \cos 360^\circ + i\sin 360^\circ = 1 + 0i = 1.$$

Итак, $\alpha^1 = \alpha$ есть корень уравнения $x^n = 1$. То же справедливо относительно следующей вершины

$$\alpha^2 = \cos\frac{720^\circ}{n} + i\sin\frac{720^\circ}{n}.$$

Мы убедимся в этом, если напишем

$$(\alpha^2)^n = \alpha^{2n} = (\alpha^n)^2 = 1^2 = 1,$$

или же воспользуемся формулой Муавра

$$(\alpha^2)^n = \cos\left(n \cdot \frac{720^\circ}{n}\right) + i\sin\left(n \cdot \frac{720^\circ}{n}\right) = \cos 720^\circ + i\sin 720^\circ = 1 + 0i = 1.$$

Точно так же мы заключаем, что все n чисел

$$1,\ \alpha,\ \alpha^2,\ \alpha^3,\ \ldots,\ \alpha^{n-1}$$

являются корнями степени n из 1. Если будем степени увеличивать дальше или рассмотрим отрицательные степени, то новых корней не получим. В самом деле,

$$\alpha^{-1} = \frac{1}{\alpha} = \frac{\alpha^n}{\alpha} = \alpha^{n-1};$$

точно так же

$$\alpha^n = 1, \quad \alpha^{n+1} = (\alpha^n)\,\alpha = 1 \cdot \alpha = \alpha,$$

и т. д., так что ранее полученные корни повторяются. Читателю предоставляем в качестве упражнения показать, что иных корней, кроме перечисленных, рассматриваемое уравнение не имеет.

Если n четное, то одна из вершин n-угольника попадает в точку -1, в соответствии с общеизвестным алгебраическим фактом: -1 есть корень четной степени из 1.

Уравнение, которому удовлетворяют корни n-й степени из 1,

$$x^n - 1 = 0, \tag{13}$$

есть уравнение n-й степени, но легко понизить его степень на единицу. Воспользуемся алгебраической формулой

$$(x^n - 1) = (x-1)(x^{n-1} + x^{n-2} + x^{n-3} + \ldots + 1). \tag{14}$$

Так как произведение двух чисел равно 0 в том и только том случае, если один из множителей равен нулю, то выражение (14) обращается в нуль или при $x = 1$, или при условии, что удовлетворяется уравнение

$$x^{n-1} + x^{n-2} + x^{n-3} + \ldots + x + 1 = 0. \tag{15}$$

Этому уравнению удовлетворяют корни α, α^2, …, α^{n-1}; оно называется *циклотомическим*, или *уравнением деления окружности*. Так, например, мнимые кубические корни из 1

$$\alpha = \cos 120^\circ + i\sin 120^\circ = \frac{-1 + i\sqrt{3}}{2},$$

$$\alpha^2 = \cos 240^\circ + i\sin 240^\circ = \frac{-1 - i\sqrt{3}}{2}$$

являются корнями уравнения

$$x^2 + x + 1 = 0,$$

как читатель сможет убедиться, выполняя подстановки. Таким же образом корни пятой степени из 1 (кроме самого числа 1) удовлетворяют уравнению

$$x^4 + x^3 + x^2 + x + 1 = 0. \tag{16}$$

Чтобы построить правильный пятиугольник, нам приходится решить уравнение четвертой степени. Простое алгебраическое ухищрение — замена $w = x + \dfrac{1}{x}$ — приводит к уравнению второй степени. Мы делим уравнение (16) на x^2 и переставляем члены:

$$x^2 + \frac{1}{x^2} + x + \frac{1}{x} + 1 = 0,$$

и, принимая во внимание, что $\left(x + \dfrac{1}{x}\right)^2 = x^2 + \dfrac{1}{x^2} + 2$, получаем

$$w^2 + w - 1 = 0.$$

По формуле (7) пункта 1 корни этого квадратного уравнения имеют вид

$$w_1 = \frac{-1 + \sqrt{5}}{2}, \qquad w_2 = \frac{-1 - \sqrt{5}}{2}.$$

Итак, мнимые корни пятой степени из 1 являются корнями следующих двух квадратных уравнений:

$$x + \frac{1}{x} = w_1, \quad \text{или} \quad x^2 + \frac{1}{2}(\sqrt{5} - 1)\,x + 1 = 0,$$

и

$$x + \frac{1}{x} = w_2, \quad \text{или} \quad x^2 - \frac{1}{2}(\sqrt{5} + 1)\,x + 1 = 0.$$

Читатель сможет их решить по той же формуле (7).

Упражнения. 1) Найдите корни 6-й степени из 1.

2) Вычислите $(1 + i)^{11}$.

3) Вычислите все различные значения выражений

$$\sqrt{1 + i}, \qquad \sqrt[3]{7 - 4i}, \qquad \sqrt[3]{i}, \qquad \sqrt[5]{-i}.$$

4) Вычислите $\dfrac{1}{2i}(i^7 - i^{-7})$.

***4. Основная теорема алгебры.** Не только уравнения вида $ax^2 + bx + c = 0$ или $x^n - 1 = 0$ разрешимы в поле комплексных чисел, но можно утверждать гораздо больше: *всякое алгебраическое уравнение степени n с действительными или комплексными коэффициентами*

$$f(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_1 x + a_0 = 0 \tag{17}$$

разрешимо в поле комплексных чисел. Для случая уравнений 3-й и 4-й степеней эта теорема была установлена в XVI в. Тартальей, Кардано и другими: оказалось, что такие уравнения решаются посредством формул, подобных формуле квадратного уравнения, но значительно более сложных. В течение почти двух столетий длилось настойчивое изучение

общего уравнения 5-й и более высоких степеней, но все усилия разрешить их теми же методами оказались напрасными. Когда молодому Гауссу в его докторской диссертации (1799) удалось впервые доказать, что решения *существуют*, то это уже было крупнейшим успехом; правда, вопрос о возможности обобщить на случай степеней $\geqslant 5$ классические формулы, позволяющие находить корни с помощью рациональных операций и извлечения корней, оставался в то время открытым (см. стр. 138).

Теорема Гаусса утверждает, что, *каково бы ни было алгебраическое уравнение вида* (17), *где n — целое положительное число, а коэффициенты a — действительные или даже комплексные числа, существует по крайней мере одно такое комплексное число* $\alpha = c + di$, *что*

$$f(a) = 0.$$

Число α называется *корнем* уравнения (17). Доказательство этой теоремы будет приведено в этой книге на стр. 289–291. Предположим пока, что теорема доказана, и выведем из нее другую теорему, известную под названием *основной теоремы алгебры* (было бы, впрочем, правильнее назвать ее основной теоремой комплексной числовой системы): *всякий алгебраический полином степени n*

$$f(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0 \tag{18}$$

может быть представлен в виде произведения ровно n множителей:

$$f(x) = (x - \alpha_1)(x - \alpha_2)\ldots(x - \alpha_n), \tag{19}$$

где $\alpha_1, \alpha_2, \ldots, \alpha_n$ — комплексные числа, корни уравнения $f(x) = 0$. Так, например, полином

$$f(x) = x^4 - 1$$

разлагается на множители следующим образом:

$$f(x) = (x - 1)(x - i)(x + i)(x + 1).$$

Что числа α являются корнями уравнения $f(x) = 0$, это очевидно из самого разложения (19), так как при $x = \alpha_r$ один из множителей $f(x)$, а следовательно, и сам полином $f(x)$, обращается в нуль.

В иных случаях не все множители $x - \alpha_1$, $x - \alpha_2$, … полинома $f(x)$ степени n оказываются различными; так, в примере

$$f(x) = x^2 - 2x + 1 = (x - 1)(x - 1)$$

мы имеем только один корень, $x = 1$, «считаемый дважды», или «кратности 2». Во всяком случае, полином степени n не может разлагаться в произведение более чем n различных множителей вида $x - \alpha$, и соответствующее уравнение не может иметь более n корней.

При доказательстве основной теоремы алгебры мы воспользуемся — не в первый раз — алгебраическим тождеством

$$x^k - \alpha^k = (x - \alpha)(x^{k-1} + \alpha x^{k-2} + \alpha^2 x^{k-3} + \ldots + \alpha^{k-2}x + \alpha^{k-1}), \tag{20}$$

которое при $\alpha = 1$ служило нам для определения суммы геометрической прогрессии. Предполагая теорему Гаусса доказанной, допустим, что $\alpha = \alpha_1$ есть корень уравнения (17), так что

$$f(\alpha_1) = \alpha_1^n + a_{n-1}\alpha_1^{n-1} + a_{n-2}\alpha_1^{n-2} + \ldots + a_1\alpha^1 + a_0 = 0.$$

Вычитая это выражение из $f(x)$ и перегруппировывая члены, мы получим тождество

$$f(x) = f(x) - f(\alpha_1) = (x^n - \alpha_1^n) + a_{n-1}(x^{n-1} - \alpha_1^{n-1}) + \ldots + a_1(x - \alpha_1). \tag{21}$$

Пользуясь теперь формулой (20), мы можем выделить множитель $x - \alpha_1$ из каждого члена и затем вынести его за скобку, причем степень многочлена, остающегося в скобках, станет уже на единицу меньше. Перегруппировывая снова члены, мы получим тождество

$$f(x) = (x - \alpha_1)g(x),$$

где $g(x)$ — многочлен степени $n - 1$:

$$g(x) = x^{n-1} + b_{n-2}x^{n-2} + \ldots + b_1x + b_0.$$

(Вычисление коэффициентов, обозначенных через b, нас здесь не интересует.) Применим дальше то же рассуждение к многочлену $g(x)$. По теореме Гаусса, существует корень α_2 уравнения $g(x) = 0$, так что

$$g(x) = (x - \alpha_2)h(x),$$

где $h(x)$ — новый многочлен степени уже $n - 2$. Повторяя эти рассуждения $n - 1$ раз (подразумевается, конечно, применение принципа математической индукции), мы, в конце концов, приходим к разложению

$$f(x) = (x - \alpha_1)(x - \alpha_2)\ldots(x - \alpha_n). \tag{22}$$

Из тождества (22) следует не только то, что комплексные числа α_1, α_2, ..., α_n суть корни уравнения (17), но и то, что иных корней уравнение (17) не имеет. Действительно, если бы число y было корнем уравнения (17), то из (22) следовало бы

$$f(y) = (y - \alpha_1)(y - \alpha_2)\ldots(y - \alpha_n) = 0.$$

Но мы видели (стр. 115), что произведение комплексных чисел равно нулю в том и *только том* случае, если один из множителей равен нулю. Итак, один из множителей $y - \alpha_r$ равен 0, т. е. $y = \alpha_r$, что и требовалось установить.

§ 6. Алгебраические и трансцендентные числа

1. Определение и вопросы существования. *Алгебраическим числом* называется всякое число x, действительное или мнимое, удовлетворяющее некоторому алгебраическому уравнению вида

$$a_nx^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0 = 0 \qquad (n \geqslant 1,\ a_n \neq 0), \tag{1}$$

где числа a_i целые. Так, например, число $\sqrt{2}$ алгебраическое, так как оно удовлетворяет уравнению

$$x^2 - 2 = 0.$$

Таким же образом алгебраическим числом является всякий корень любого уравнения с целыми коэффициентами третьей, четвертой, пятой, какой угодно степени, и независимо от того, выражается или не выражается он в радикалах. Понятие алгебраического числа есть естественное обобщение понятия рационального числа, которое соответствует частному случаю $n = 1$.

Не всякое действительное число является алгебраическим. Это вытекает из следующей, высказанной Кантором, теоремы: *множество всех алгебраических чисел счетно.* Так как множество всех действительных чисел несчетное, то обязательно должны существовать действительные числа, не являющиеся алгебраическими.

Укажем один из методов пересчета множества алгебраических чисел. Каждому уравнению вида (1) сопоставим целое положительное число

$$h = |a_n| + |a_{n-1}| + \ldots + |a_1| + |a_0| + n,$$

которое назовем ради краткости «высотой» уравнения. Для каждого *фиксированного* значения n существует лишь *конечное* число уравнений вида (1) с высотой h. Каждое из таких уравнений имеет самое большее n корней. Поэтому может существовать лишь конечное число алгебраических чисел, порождаемых уравнениями с высотой h; следовательно, все алгебраические числа можно расположить в виде последовательности, перечисляя сначала те из них, которые порождаются уравнениями высоты 1, затем — высоты 2 и т. д.

Это доказательство счетности множества алгебраических чисел устанавливает существование действительных чисел, которые не являются алгебраическими. Такие числа называют *трансцендентными* (от латинского *transcendere* — переходить, превосходить); такое наименование им дал Эйлер, потому что они «превосходят мощность алгебраических методов».

Канторово доказательство существования трансцендентных чисел не принадлежит к числу конструктивных. Теоретически рассуждая, можно было бы построить трансцендентное число с помощью диагональной процедуры, производимой над воображаемым списком десятичных разложений всех алгебраических чисел; но такая процедура лишена всякого практического значения и не привела бы к числу, разложение которого в десятичную (или какую-нибудь иную) дробь можно было бы на самом деле написать. Наиболее интересные проблемы, связанные с трансцендентными числами, заключаются в доказательстве того, что определенные, конкретные числа (сюда относятся числа π и e, о которых см. стр. 319–322) являются трансцендентными.

****2. Теорема Лиувилля и конструирование трансцендентных чисел.** Доказательство существования трансцендентных чисел еще до Кантора было дано Ж. Лиувиллем (1809–1862). Оно дает возможность на самом деле конструировать примеры таких чисел. Доказательство Лиувилля более трудно, чем доказательство Кантора, и это неудивительно, так как сконструировать пример, вообще говоря, сложнее, чем доказать существование. Приводя ниже доказательство Лиувилля, мы имеем в виду только подготовленного читателя, хотя для понимания доказательства совершенно достаточно знания элементарной математики.

Как обнаружил Лиувилль, иррациональные алгебраические числа обладают тем свойством, что они не могут быть приближены рациональными числами с очень большой степенью точности, если только не взять знаменатели приближающих дробей чрезвычайно большими.

Предположим, что число z удовлетворяет алгебраическому уравнению с целыми коэффициентами

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n = 0 \qquad (a_n \neq 0), \tag{2}$$

но не удовлетворяет такому же уравнению более низкой степени. Тогда говорят, что само x есть алгебраическое число *степени* n. Так, например, число $z = \sqrt{2}$ есть алгебраическое число степени 2, так как удовлетворяет уравнению $x^2 - 2 = 0$ степени 2, но не удовлетворяет уравнению первой степени; число $z = \sqrt[3]{2}$ — степени 3, так как удовлетворяет уравнению $x^3 - 2 = 0$, но не удовлетворяет (как мы покажем в главе III) уравнению более низкой степени. Алгебраическое число степени $n > 1$ не может быть рациональным, так как рациональное число $z = \frac{p}{q}$ удовлетворяет уравнению $qx - p = 0$ степени 1. Каждое иррациональное число z может быть с какой угодно степенью точности приближено с помощью рационального числа; это означает, что всегда можно указать последовательность рациональных чисел

$$\frac{p_1}{q_1}, \ \frac{p_2}{q_2}, \ \ldots$$

с неограниченно растущими знаменателями, обладающую тем свойством, что

$$\frac{p_r}{q_r} \to z.$$

Теорема Лиувилля утверждает: каково бы ни было алгебраическое число z степени $n > 1$, оно не может быть приближено посредством рационального числа $\frac{p}{q}$ с точностью лучшей чем $\frac{1}{q^{n+1}}$; другими словами, при достаточно больших знаменателях непременно выполняется неравенство

$$\left| z - \frac{p}{q} \right| > \frac{1}{q^{n+1}}. \tag{3}$$

Мы собираемся привести доказательство этой теоремы, но раньше покажем, как с ее помощью можно строить трансцендентные числа. Рассмотрим число

$$z = a_1 \cdot 10^{-1!} + a_2 \cdot 10^{-2!} + a_3 \cdot 10^{-3!} + \ldots + a_m \cdot 10^{-m!} + \ldots =$$
$$= 0{,}a_1 a_2 000 a_3 00000000000000000 a_4 000\ldots,$$

где a_i обозначают произвольные цифры от 1 до 9 (проще всего было бы положить все a_i равными 1), а символ $n!$, как обычно (см. стр. 36), обозначает $1 \cdot 2 \cdot \ldots \cdot n$. Характерным свойством десятичного разложения такого числа является то, что быстро возрастающие по своей длине группы нулей чередуются в нем с отдельными цифрами, отличными от нуля. Обозначим через z_m конечную десятичную дробь, получающуюся, когда в разложении возьмем все члены до $a_m \cdot 10^{-m!}$ включительно. Тогда получим неравенство

$$|z - z_m| < 10 \cdot 10^{-(m+1)!}. \tag{4}$$

Предположим, что z было бы алгебраическим числом степени n. Тогда, полагая в неравенстве Лиувилля (3) $\frac{p}{q} = z_m = \frac{p}{10^{m!}}$, мы должны иметь

$$|z - z_m| > \frac{1}{10^{(n+1)m!}}$$

при достаточно больших значениях m. Сопоставление последнего неравенства с неравенством (4) дает

$$\frac{1}{10^{(n+1)m!}} < \frac{10}{10^{(m+1)!}} = \frac{1}{10^{(m+1)!-1}},$$

откуда следует $(n+1)m! > (m+1)! - 1$ при достаточно больших m. Но это неверно для значений m, больших чем n (пусть читатель потрудится дать детализированное доказательство этого утверждения). Мы пришли к противоречию. Итак, число z — трансцендентное.

Остается доказать теорему Лиувилля. Предположим, что z — алгебраическое число степени $n > 1$, удовлетворяющее уравнению (1), так что

$$f(z) = 0. \tag{5}$$

Пусть $z_m = \frac{p_m}{q_m}$ — последовательность рациональных чисел, причем $z_m \to z$. Тогда

$$f(z_m) = f(z_m) - f(z) = a_1(z_m - z) + a_2(z_m^2 - z^2) + \ldots + a_n(z_m^n - z^n).$$

Деля обе части на $z_m - z$ и пользуясь алгебраической формулой

$$\frac{u^n - v^n}{u - v} = u^{n-1} + u^{n-2}v + u^{n-3}v^2 + \ldots + uv^{n-2} + v^{n-1},$$

мы получаем:

$$\frac{f(z_m)}{z_m - z} = a_1 + a_2(z_m + z) + a_3(z_m^2 + z_m z + z^2) + \ldots$$
$$\ldots + a_n(z_m^{n-1} + \ldots + z^{n-1}). \tag{6}$$

Так как z_m стремится к z, то при достаточно больших m рациональное число z_m будет отличаться от z меньше чем на единицу. Поэтому при достаточно больших m можно сделать следующую грубую оценку:

$$\left|\frac{f(z_m)}{z_m - z}\right| < |a_1| + 2|a_2|(|z|+1) + 3|a_3|(|z|+1)^2 + \ldots$$
$$\ldots + n|a_n|(|z|+1)^{n-1} = M, \quad (7)$$

причем стоящее справа число M — постоянное, так как z не меняется в процессе доказательства. Выберем теперь m настолько большим, чтобы у дроби $z_m = \dfrac{p_m}{q_m}$ знаменатель q_m был больше, чем M; тогда

$$|z - z_m| > \frac{|f(z_m)|}{M} > \frac{|f(z_m)|}{q_m}. \quad (8)$$

Ради краткости условимся дальше писать p вместо p_m и q вместо q_m. В таком случае

$$|f(z_m)| = \left|\frac{a_0 q^n + a_1 q^{n-1} p + \ldots + a_n p^n}{q^n}\right|. \quad (9)$$

Рациональное число $z_m = \dfrac{p}{q}$ не может быть корнем уравнения $f(x) = 0$, так как тогда можно было бы из многочлена $f(x)$ выделить множитель $(x - z_m)$, и, значит, z удовлетворяло бы уравнению степени низшей чем n. Итак, $f(z_m) \neq 0$. Но числитель в правой части равенства (9) есть целое число и, следовательно, по абсолютной величине он по меньшей мере равен единице. Таким образом, из сопоставления соотношений (8) и (9) вытекает неравенство

$$|z - z_m| > \frac{1}{q}\,\frac{1}{q^n} = \frac{1}{q^{n+1}}, \quad (10)$$

как раз и составляющее содержание указываемой теоремы.

На протяжении нескольких последних десятилетий исследования, касающиеся возможности приближения алгебраических чисел рациональными, продвинулись гораздо дальше. Например, норвежский математик А. Туэ (1863–1922) установил, что в неравенстве Лиувилля (3) показатель $n+1$ может быть заменен меньшим показателем $\dfrac{n}{2} + 1$. К. Л. Зигель показал, что можно взять и еще меньший (еще меньший при бóльших n) показатель $2\sqrt{n}$.

Трансцендентные числа всегда были темой, приковывающей к себе внимание математиков. Но до сравнительно недавнего времени среди чисел, которые интересны сами по себе, было известно очень немного таких, трансцендентный характер которых был бы установлен. (Из трансцендентности числа π, о которой пойдет речь в главе III, следует невозможность квадратуры круга с помощью линейки и циркуля.) В своем выступлении на Парижском международном математическом конгрессе 1900 г. Давид Гильберт предложил тридцать математических

проблем, допускающих простую формулировку, некоторые — даже совсем элементарную и популярную, из которых ни одна не только не была решена, но даже и не казалась способной быть разрешенной средствами математики той эпохи. Эти «проблемы Гильберта» оказали сильное возбуждающее влияние на протяжении всего последующего периода развития математики. Почти все они мало-помалу были разрешены, и во многих случаях их решение было связано с ясно выраженными успехами в смысле выработки более общих и более глубоких методов. Одна из проблем, казавшаяся довольно безнадежной, заключалась в доказательстве того, что число

$$2^{\sqrt{2}}$$

является трансцендентным (или хотя бы иррациональным). На протяжении трех десятилетий не было даже намека на такой подход к вопросу с чьей-нибудь стороны, который открывал бы надежду на успех. Наконец, Зигель и, независимо от него, молодой русский математик А. Гельфонд открыли новые методы для доказательства трансцендентности многих чисел, имеющих значение в математике. В частности, была установлена трансцендентность не только гильбертова числа $2^{\sqrt{2}}$, но и целого довольно обширного класса чисел вида a^b, где a — алгебраическое число, отличное от 0 и 1, а b — иррациональное алгебраическое число.

ДОПОЛНЕНИЕ К ГЛАВЕ II

Алгебра множеств

1. Общая теория. Понятие *класса*, или *совокупности*, или *множества* объектов — одно из самых фундаментальных в математике. Множество определяется некоторым свойством («атрибутом») $\mathfrak{A}$, которым должен или обладать, или не обладать каждый рассматриваемый объект; те объекты, которые обладают свойством $\mathfrak{A}$, образуют множество A. Так, если мы рассматриваем целые числа и свойство $\mathfrak{A}$ заключается в том, чтобы «быть простым», то соответствующее множество A состоит из всех простых чисел 2, 3, 5, 7, ...

Математическая теория множеств исходит из того, что из множеств с помощью определенных операций можно образовывать новые множества (подобно тому как из чисел посредством операций сложения и умножения получаются новые числа). Изучение операций над множествами составляет предмет «алгебры множеств», которая имеет много общего с обыкновенной числовой алгеброй, хотя кое в чем и отличается от нее. Тот факт, что алгебраические методы могут быть применены к изучению нечисловых объектов, каковыми являются множества, иллю-

стрирует большую общность идей современной математики. В последнее время выяснилось, что алгебра множеств бросает новый свет на многие области математики, например, теорию меры и теорию вероятностей; она полезна также при систематизации математических понятий и выяснении их логических связей.

В дальнейшем I будет обозначать некоторое постоянное множество объектов, природа которых безразлична, и которое мы можем называть *универсальным множеством* (или *универсумом рассуждения*), а A, B, C, ... будут какие-то подмножества I. Если I есть совокупность всех натуральных чисел, то A, скажем, может обозначать множество всех четных чисел, B — множество всех нечетных чисел, C — множество всех простых чисел, и т. п. Если I обозначает совокупность всех точек на плоскости, то A может быть множеством точек внутри какого-то круга, B — множеством точек внутри другого круга и т. п. В число «подмножеств» нам удобно включить само I, а также «пустое» множество $\varnothing$, не содержащее никаких элементов. Цель, которую преследует такое искусственное расширение, заключается в сохранении того положения, что каждому свойству $\mathfrak{A}$ соответствует некоторое множество элементов из I, обладающих этим свойством. В случае, если $\mathfrak{A}$ есть универсально выполняемое свойство, примером которого может служить (если речь идет о числах) свойство удовлетворять тривиальному равенству $x = x$, то соответствующее подмножество I будет само I, так как каждый элемент обладает таким свойством; с другой стороны, если $\mathfrak{A}$ есть какое-то внутренне противоречивое свойство (вроде $x \neq x$), то соответствующее подмножество не содержит вовсе элементов, оно — «пустое» и обозначается символом $\varnothing$.

Говорят, что множество A есть *подмножество* множества B, короче, «A входит в B», или «B содержит A», если во множестве A нет такого элемента, который не был бы также во множестве B. Этому соотношению соответствует запись

$$A \subset B, \quad \text{или} \quad B \supset A.$$

Например, множество A всех целых чисел, делящихся на 10, есть подмножество множества B всех целых чисел, делящихся на 5, так как каждое число, делящееся на 10, делится также на 5. Соотношение $A \subset B$ не исключает соотношения $B \subset A$. Если имеет место и то и другое, то мы пишем

$$A = B.$$

Это означает, что каждый элемент A есть вместе с тем элемент B, и обратно, так что множества A и B содержат как раз одни и те же элементы.

Соотношение $A \subset B$ между множествами во многих отношениях напоминает соотношение $a \leqslant b$ между числами. В частности, отметим сле-

дующие свойства этого соотношения:

1) $A \subset A$.
2) Если $A \subset B$ и $B \subset A$, то $A = B$.
3) Если $A \subset B$ и $B \subset C$, то $A \subset C$.

По этой причине соотношение $A \subset B$ иногда называют «отношением порядка». Главное отличие рассматриваемого соотношения от соотношения $a \leqslant b$ между числами заключается в том, что между *всякими* двумя заданными (действительными) числами a и b непременно осуществляется по меньшей мере одно из соотношений $a \leqslant b$ или $b \leqslant a$, тогда как для соотношения $A \subset B$ между множествами аналогичное утверждение неверно. Например, если A есть множество, состоящее из чисел 1, 2, 3,

$$A = \{1, 2, 3\},$$

а B — множество, состоящее из чисел 2, 3, 4,

$$B = \{2, 3, 4\},$$

то не имеет места ни соотношение $A \subset B$, ни соотношение $B \subset A$. По этой причине говорят, что подмножества A, B, C, ... множества I являются «частично упорядоченными», тогда как действительные числа a, b, c, ... образуют «вполне упорядоченную» совокупность.

Заметим, между прочим, что из определения соотношения $A \subset B$ следует, что, каково бы ни было подмножество A множества I,

4) $\varnothing \subset A$

и

5) $A \subset I$.

Свойство 4) может показаться несколько парадоксальным, но, если вдуматься, оно логически строго соответствует точному смыслу определения знака $\subset$. В самом деле, соотношение $\varnothing \subset A$ нарушалось бы только в том случае, если бы пустое множество $\varnothing$ содержало элемент, который не содержался бы в A; но так как пустое множество не содержит вовсе элементов, то этого быть не может, каково бы ни было A.

Мы определим теперь две операции над множествами, формально обладающие многими алгебраическими свойствами сложения и умножения чисел, хотя по своему внутреннему содержанию совершенно отличные от этих арифметических действий. Пусть A и B — какие-то два множества. Под *объединением*, или «логической суммой», A и B понимается множество, состоящее из тех элементов, которые содержатся *или* в A, *или* в B (включая и те элементы, которые содержатся *и* в A *и* в B). Это множество обозначается $A + B$.[1] Под «пересечением», или «логическим произведением», A и B понимается множество, состоящее из тех элементов, которые содержатся *и* в A *и* в B. Это множество обозначается AB.[2]

[1] Или $A \cup B$. — *Прим. ред.*

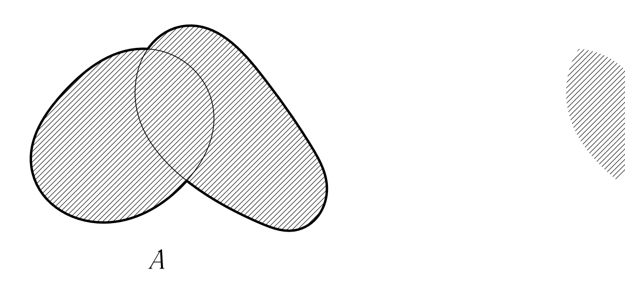


Рис. 26. Объединение и пересечение множеств

Проиллюстрируем приведенные определения примером. Возьмем опять в качестве A и B множества

$$A = \{1,\ 2,\ 3\}, \qquad B = \{2,\ 3,\ 4\}.$$

Тогда

$$A + B = \{1,\ 2,\ 3,\ 4\}, \qquad AB = \{2,\ 3\}.$$

Среди важных алгебраических свойств операций $A + B$ и AB перечислим следующие. Читатель сможет проверить их справедливость, исходя из определения самих операций:

6) $A + B = B + A$.
7) $AB = BA$.
8) $A + (B + C) = (A + B) + C$.
9) $A(BC) = (AB)C$.
10) $A + A = A$.
11) $AA = A$.
12) $A(B + C) = AB + AC$.
13) $A + (BC) = (A + B)(A + C)$.
14) $A + \varnothing = A$.
15) $AI = A$.
16) $A + I = I$.
17) $A\varnothing = \varnothing$.
18) Соотношение $A \subset B$ эквивалентно каждому из двух соотношений

$$A + B = B, \qquad AB = A.$$

Проверка всех этих законов — дело самой элементарной логики. Например, правило 10) констатирует, что множество элементов, содержащихся или в A, или в A, есть как раз множество A; правило 12) констатирует, что множество тех элементов, которые содержатся в A и вместе с тем содержатся или в B, или в C, совпадает со множеством элементов, которые или содержатся одновременно в A и в B, или содержатся одновременно в A и в C. Логические рассуждения, используемые при доказательствах подобного рода правил, удобно иллюстрируются, если мы условимся изображать множества A, B, C, … в виде некоторых фигур на плоскости и будем очень внимательны в том отношении, чтобы не упустить ни одной из возникающих логических возможностей, когда речь идет о наличии общих элементов двух множеств или, напротив, наличии в одном множестве элементов, которые не содержатся в другом.

[2] Или $A \cap B$. — *Прим. ред.*

Читатель, несомненно, обратил внимание на то обстоятельство, что законы 6), 7), 8), 9) и 12) внешне тождественны с хорошо знакомыми коммутативным, ассоциативным и дистрибутивным законами обыкновенной алгебры. Отсюда следует, что все правила обыкновенной алгебры, вытекающие из этих законов, действительны также в алгебре множеств. Напротив, законы 10), 11) и 13) не имеют своих аналогов в обыкновенной алгебре, и они придают алгебре множеств более простую структуру. Например, формула бинома в алгебре множеств сводится к простейшему равенству

$$(A+B)^n = (A+B)\cdot(A+B)\ldots(A+B) = A+B,$$

которое следует из закона 11). Законы 14), 15) и 17) говорят о том, что свойства множеств $\varnothing$ и I по отношению к операциям объединения и пересечения множеств весьма похожи на свойства чисел 0 и 1 по отношению к операциям числовых действий сложения и умножения. Но закон 16) не имеет аналога в числовой алгебре.

Остается дать определение еще одной операции в алгебре множеств. Пусть A — какое-нибудь подмножество универсального множества I. Тогда под *дополнением* A в I понимается множество всех элементов I, которые *не* содержатся в A. Для этого множества мы введем обозначение A'. Так, если I есть множество всех натуральных чисел, а A — множество всех простых чисел, то A' есть множество, состоящее из всех составных чисел и числа 1. Операция перехода от A к A', для которой нет аналога в обыкновенной алгебре, обладает следующими свойствами:

19) $A + A' = I$. 20) $AA' = \varnothing$.
21) $\varnothing' = I$. 22) $I' = \varnothing$.
23) $A'' = A$.
24) Соотношение $A \subset B$ эквивалентно соотношению $B' \subset A'$.
25) $(A+B)' = A'B'$. 26) $(AB)' = A' + B'$.

Проверку этих свойств мы опять предоставляем читателю.

Законы 1)–26) лежат в основе алгебры множеств. Они обладают замечательным свойством «двойственности» в следующем смысле:

Если в одном из законов 1)–26) *заменить друг на друга соответственно символы*

$$\subset \text{ и } \supset$$
$$\varnothing \text{ и } I$$
$$+ \text{ и } \cdot$$

(*в каждом их вхождении*), *то в результате снова получается один из этих же законов.* Например, закон 6) переходит в закон 7), 12) — в 13), 17) — в 16) и т. д. Отсюда следует, что *каждой теореме, которая может быть выведена из законов* 1)–26), *соответствует другая, «двойственная» ей теорема, получающаяся из первой посредством указанных перестановок символов.* В самом деле, так как доказательство

первой теоремы состоит из последовательного применения (на различных стадиях проводимого рассуждения) некоторых из законов 1–26), то применение на соответствующих стадиях «двойственных» законов составит доказательство «двойственной» теоремы. (По поводу подобной же «двойственности» в геометрии см. главу IV.)

2. Применение к математической логике. Проверка законов алгебры множеств основывалась на анализе логического смысла соотношения $A \subset B$ и операций $A + B$, AB и A'. Мы можем теперь обратить этот процесс и рассматривать законы 1)–26) как базу для «алгебры логики». Скажем точнее: та часть логики, которая касается множеств, или, что по существу то же, свойств рассматриваемых объектов, может быть сведена к формальной алгебраической системе, основанной на законах 1)–26). Логическая «условная вселенная» определяет множество I; *каждое свойство* $\mathfrak{A}$ *определяет множество* A, *состоящее из тех объектов в* I, *которые обладают этим свойством.* Правила перевода обычной логической терминологии на язык множеств ясны из следующих примеров:

«A или B»	$A + B$
«A и B»	AB
«Не A»	A'
«Ни A, ни B»	$(A + B)'$, или, что то же, $A'B'$
«Неверно, что и A, и B»	$(AB)'$, или, что то же, $A' + B'$
«Всякое A есть B», или «Если A, то B», или «Из A следует B»	$A \subset B$
«Какое-то A есть B»	$AB \neq \varnothing$
«Никакое A не есть B»	$AB = \varnothing$
«Какое-то A не есть B»	$AB' \neq \varnothing$
«Нет никакого A»	$A = \varnothing$

В терминах алгебры множеств силлогизм «Barbara», обозначающий, что «если всякое A есть B и всякое B есть C, то всякое A есть C», принимает простой вид:

3) Если $A \subset B$ и $B \subset C$, то $A \subset C$.

Аналогично «закон противоречия», утверждающий, что «объект не может одновременно обладать и не обладать некоторым свойством», записывается в виде:

20) $AA' = \varnothing$,

а «закон исключенного третьего», говорящий, что «объект должен или обладать, или не обладать некоторым свойством», записывается:

19) $A + A' = I$.

Таким образом, та часть логики, которая выразима в терминах символов $\subset$, $+$, $\cdot$ и $'$, может трактоваться как формальная алгебраическая система, подчиненная законам 1)–26). На основе слияния логического анализа математики и математического анализа логики создалась новая дисциплина — *математическая логика*, которая в настоящее время находится в процессе бурного развития.

С аксиоматической точки зрения заслуживает внимания тот замечательный факт, что утверждения 1)–26), вместе со всеми прочими теоремами алгебры множеств, могут быть логически выведены из следующих трех равенств:

27) $A + B = B + A$,
$(A + B) + C = A + (B + C)$,
$(A' + B')' + (A' + B)' = A$.

Отсюда следует, что алгебра множеств может быть построена как чисто дедуктивная теория, вроде евклидовой геометрии, на базе этих трех положений, принимаемых в качестве аксиом. Если эти аксиомы приняты, то операция AB и отношение $A \subset B$ *определяются* в терминах $A + B$ и A':

$$AB \text{ обозначает множество } (A' + B')',$$
$$A \subset B \text{ обозначает, что } A + B = B.$$

Совершенно иного рода пример математической системы, в которой выполняются все формальные законы алгебры множеств, дается системой восьми чисел 1, 2, 3, 5, 6, 10, 15, 30: здесь $a + b$ обозначает, по определению, общее наименьшее кратное a и b, ab — общий наибольший делитель a и b, $a \subset b$ — утверждение «b делится на a» и a' — число $\frac{30}{a}$. Существование таких примеров повлекло за собой изучение общих алгебраических систем, удовлетворяющих законам 27). Такие системы называются «булевыми алгебрами» — в честь Джорджа Буля (1815–1864), английского математика и логика, книга которого «An investigation of the laws of thought» (Исследование законов мышления) появилась в 1854 г.

3. Одно из применений к теории вероятностей. Алгебра множеств имеет ближайшее отношение к теории вероятностей и позволяет взглянуть на нее в новом свете. Рассмотрим простейший пример: представим себе эксперимент с конечным числом возможных исходов, которые все мыслятся как «равновозможные». Эксперимент может, например, заключаться в том, что мы вытягиваем наугад карту из хорошо перетасованной полной колоды. Если множество всех исходов эксперимента обозначим через I, а A обозначает какое-нибудь подмножество I, то вероятность того, что исход эксперимента окажется принадлежащим к подмножеству A, определяется как отношение

$$p(A) = \frac{\textit{число элементов } A}{\textit{число элементов } I}.$$

Если условимся число элементов в каком-нибудь множестве A обозначать через $n(A)$, то последнему равенству можно придать вид

$$p(A) = \frac{n(A)}{n(I)}. \tag{1}$$

В нашем примере, допуская, что A есть подмножество треф, мы получим $n(A) = 13$, $n(I) = 52$ и $p(A) = \frac{13}{52} = \frac{1}{4}$.

Идеи алгебры множеств обнаруживаются при вычислении вероятностей тогда, когда приходится, зная вероятности одних множеств, вычислять вероятности других. Например, зная вероятности $p(A)$, $p(B)$ и $p(AB)$, можно вычислить вероятность $p(A + B)$:

$$p(A + B) = p(A) + p(B) - p(AB). \tag{2}$$

Доказать это не составит труда. Мы имеем

$$n(A + B) = n(A) + n(B) - n(AB),$$

так как элементы, содержащиеся одновременно в A и в B, т. е. элементы AB, считаются дважды при вычислении суммы $n(A) + n(B)$, и, значит, нужно вычесть $n(AB)$ из этой суммы, чтобы подсчет $n(A + B)$ был произведен правильно. Деля затем обе части равенства на $n(I)$, мы получаем соотношение (2).

Более интересная формула получается, если речь идет о трех множествах A, B, C из I. Пользуясь соотношением (2), мы имеем

$$p(A + B + C) = p[(A + B) + C] = p(A + B) + p(C) - p[(A + B)C].$$

Закон (12) из предыдущего пункта дает нам $(A + B)C = AC + BC$. Отсюда следует:

$$p[(A + B)C)] = p(AC + BC) = p(AC) + p(BC) - p(ABC).$$

Подставляя в полученное раньше соотношение значение $p[(A + B)C]$ и значение $p(A + B)$, взятое из (2), мы приходим к нужной нам формуле:

$$p(A + B + C) = p(A) + p(B) + p(C) - p(AB) - p(AC) - p(BC) + p(ABC). \tag{3}$$

В качестве примера рассмотрим следующий эксперимент. Три цифры 1, 2, 3 пишутся в каком попало порядке. Какова вероятность того, что по крайней мере одна из цифр окажется на надлежащем (в смысле нумерации) месте? Пусть A есть множество перестановок, в которых цифра 1 стоит на первом месте, B — множество перестановок, в которых цифра 2 стоит на втором месте, C — множество перестановок, в которых цифра 3 стоит на третьем месте. Нам нужно вычислить $p(A + B + C)$. Ясно, что

$$p(A) = p(B) = p(C) = \frac{2}{6} = \frac{1}{3};$$

действительно, если какая-нибудь цифра стоит на надлежащем месте, то имеются две возможности переставить остальные две цифры из общего числа $3 \cdot 2 \cdot 1 = 6$ возможных перестановок трех цифр. Далее,

$$p(AB) = p(AC) = p(BC) = \frac{1}{6}, \qquad p(ABC) = \frac{1}{6},$$

так как в каждом из этих случаев возникает только одна возможность. И тогда формула (3) дает нам

$$p(A + B + C) = 3 \cdot \frac{1}{6} - 3 \cdot \frac{1}{6} + \frac{1}{6} = 1 - \frac{1}{2} + \frac{1}{6} = 0{,}6666\ldots$$

Упражнение. Выведите соответствующую формулу для $p(A + B + C + D)$ и примените ее к эксперименту, в котором будут участвовать 4 цифры. Соответствующая вероятность равна $\frac{5}{8} = 0{,}6250$.

Общая формула для объединения n множеств имеет вид

$$p(A_1 + A_2 + \ldots + A_n) = \\ = \sum_1 p(A_i) - \sum_2 p(A_i A_j) + \sum_3 p(A_i A_j A_k) - \ldots \pm p(A_1 A_2 \ldots A_n), \quad (4)$$

где символы $\sum\limits_1$, $\sum\limits_2$, $\sum\limits_3$, $\ldots$, $\sum\limits_{n-1}$ обозначают суммирование по всем возможным комбинациям, содержащим одну, две, три, ..., $(n-1)$ букв из числа A_1, A_2, ..., A_n. Эта формула может быть установлена посредством математической индукции — точно так же, как формула (3) была выведена из формулы (2).

Из формулы (4) можно заключить, что если n цифр $1, 2, 3, \ldots, n$ написаны в каком угодно порядке, то вероятность того, что по крайней мере одна из цифр окажется на надлежащем месте, равна

$$p_n = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \ldots \pm \frac{1}{n!}, \quad (5)$$

причем перед последним членом стоит знак $+$ или $-$, смотря по тому, является ли n четным или нечетным. В частности, при $n = 5$ эта вероятность равна

$$p_5 = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \frac{1}{5!} = \frac{19}{30} = 0{,}6333\ldots$$

В главе VIII мы увидим, что, когда n стремится к бесконечности, выражение

$$S_n = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots \pm \frac{1}{n!}$$

стремится к пределу $\frac{1}{e}$, значение которого, с пятью знаками после запятой, равно 0,36788. Так как из формулы (5) видно, что $p_n = 1 - S_n$, то отсюда следует, что при $n \to \infty$

$$p_n \to 1 - \frac{1}{e} \approx 0{,}63212.$$

ГЛАВА III

Геометрические построения. Алгебра числовых полей

Введение

Задачи на построение всегда были одним из самых любимых предметов геометрических занятий. С помощью только циркуля и линейки, как читатель знает из школьного курса, можно выполнить очень много разнообразных построений: разделить пополам отрезок или угол, провести через точку перпендикуляр к данной прямой, вписать в данный круг правильный шестиугольник и т. д. Во всех этих построениях линейка служит только для того, чтобы проводить прямую линию, но не для того, чтобы измерять или откладывать расстояния. Традиционное ограничение — пользоваться только циркулем и линейкой — восходит к глубокой древности, хотя на практике сами греки без колебания прибегали и к другим инструментам.

Одной из самых знаменитых, классических задач на построение является задача Аполлония (около 220 года до нашей эры): даны три круга, требуется провести четвертый, касательный к трем данным. В частности, не исключено, что один или большее число из данных кругов «вырождаются» в точку или прямую («круг» с «нулевым» или с «бесконечным» радиусом). Например, может идти речь о проведении круга, касательного к двум данным прямым и проходящего через данную точку. Если такого рода специальные случаи не связаны с затруднениями, то в общей постановке задача принадлежит к числу весьма трудных.

Из всех задач на построение задача построения (с помощью циркуля и линейки) правильного n-угольника представляет, может быть, наибольший интерес. Для ряда значений n, например, $n = 3$, 4, 5, 6, решение было известно уже в древности и излагается в школьной геометрии. Но в случае правильного семиугольника ($n = 7$) построение, как было доказано, невозможно. Вот еще три классические проблемы, решение которых разыскивалось долго и безрезультатно: разделить на три равные части данный произвольный угол, удвоить данный куб (т. е. построить сторону куба, объем которого вдвое больше, чем объем куба, сторона которого задана) и выполнить «квадратуру» круга (т. е. построить квадрат,

имеющий такую же площадь, как и данный круг). И в этих проблемах предполагается, что, кроме циркуля и линейки, другие инструменты не применяются.

Проблемы подобного рода, не поддающиеся решению, привели к одному из самых замечательных и оригинальных направлений математической мысли. После нескольких столетий безуспешных поисков математики утвердились в подозрении, что найти решение невозможно. На очередь встал соблазнительный по своей трудности новый вопрос: *как можно доказать, что та или иная проблема не может быть разрешена*?

В области алгебры тот же вопрос возник в связи с проблемой решения уравнений 5-й и более высоких степеней. В течение XVI столетия было установлено, что алгебраические уравнения степени 3 и 4 решаются посредством той же процедуры, что и квадратные. Эта процедура может быть, вообще говоря, охарактеризована следующим образом: решения, или «корни», уравнения представляются в виде выражений, составленных из коэффициентов уравнения и содержащих операции, из которых каждая есть или рациональная — сложение, вычитание, умножение, деление, — или же извлечение корня — квадратного, кубического или четвертой степени. Говорят короче, что алгебраическое уравнение не выше четвертой степени «решается в радикалах» (radix по-латыни означает «корень»). Казалось как нельзя более естественным пытаться обобщить эту процедуру на уравнения 5-й и более высоких степеней, пользуясь, конечно, и радикалами соответствующих степеней. Но ни одна из попыток не увенчалась успехом. В XVIII столетии были случаи, когда даже выдающиеся математики впадали в заблуждение, предполагая, что решение ими найдено. Но только в начале XIX столетия у итальянца Руффини (1765–1822) и у гениального норвежского математика Н. Г. Абеля (1802–1829) возникла поистине революционная для того времени идея — *доказать невозможность решения в радикалах общего алгебраического уравнения степени* n. Нужно понимать совершенно отчетливо, что речь не идет о *существовании* решения алгебраического уравнения степени n: существование решений было строго доказано Гауссом в его докторской диссертации в 1799 г. Таким образом, уже не было никаких сомнений в том, что каждое алгебраическое уравнение действительно имеет корни, в особенности после того, как были указаны приближенные методы для их вычисления с какой угодно степенью точности. «Численное» решение алгебраических уравнений, имеющее громадное значение в приложениях, прекрасно разработано. Проблема Абеля и Руффини была поставлена совсем иначе: *может ли быть найдено решение с помощью одних только рациональных операций и операций извлечения корней*? Именно стремление добиться полной ясности в этом вопросе послужило толчком для великолепного развития современной алгебры и теории

групп, начатого работами Руффини, Абеля и Э. Галуа (1811–1832).

Доказательство невозможности некоторых геометрических построений оказывается примером, иллюстрирующим направление в алгебре, о котором только что было сказано. Именно оперируя алгебраическими понятиями, мы сможем установить в этой главе невозможность и трисекции угла, и построения правильного семиугольника, и удвоения куба с помощью одних только циркуля и линейки. (Проблема квадратуры круга значительно сложнее; см. по этому поводу стр. 160.) Подходя ближе к интересующему нас вопросу, мы сосредоточимся не на его отрицательной стороне — невозможности выполнения тех или иных построений, а придадим ему положительный характер: как могут быть полностью охарактеризованы задачи на построение, допускающие решение? После того как ответ на этот вопрос будет найден, не составит труда установить, что рассматриваемые нами проблемы не входят в эту категорию.

В возрасте 17 лет Гаусс исследовал возможность построения правильных «p-угольников», где p — простое число. В то время были известны построения только для случаев $p = 3$ и $p = 5$. Гаусс установил, что построения возможны в том и только том случае, если p есть простое «число Ферма»:

$$p = 2^{2^n} + 1.$$

Первые числа Ферма суть 3, 5, 17, 257, 65 537 (см. стр. 44). Это открытие произвело на Гаусса такое впечатление, что он сразу отказался от филологической карьеры и решил посвятить свою жизнь математике и ее приложениям. Он и позднее смотрел на это первое из своих открытий с особенной гордостью. После смерти Гаусса в Гёттингене была воздвигнута его бронзовая статуя, с пьедесталом в форме правильного 17-угольника. Трудно придумать более достойную почесть.

Когда речь идет о геометрических построениях, никак не следует упускать из виду, что проблема заключается не в практическом вычерчивании фигур с известной степенью аккуратности, а в том, может ли построение быть выполнено теоретически, предполагая, что наши инструменты дают абсолютную точность. Гаусс доказал именно принципиальную возможность рассмотренных им построений. Его теория не касается того, как выполнить построение на самом деле, какие следует использовать приемы, чтобы упростить процедуру или даже уменьшить число необходимых конструктивных операций. Все это — вопросы не столь высокого теоретического значения. С практической точки зрения, такие построения не дают столь удовлетворительного результата, какой может быть достигнут посредством хорошего транспортира. Вероятно, именно непониманием теоретического характера вопроса о геометрических построениях, с одной стороны, а с другой — упорным нежеланием считаться с прекрасно установленными научными фактами нужно объяснять то обстоятельство, что еще продолжают существовать нескончае-

мые вереницы «трисекторов» и «квадратурщиков». Тем из них, которые способны понимать элементарную математику, можно порекомендовать заняться изучением этой главы.

В заключение отметим, что в известном отношении наша постановка вопроса о геометрических построениях представляется искусственной. Циркуль и линейка, конечно, простейшие из геометрических инструментов, но требование ограничиваться именно этими инструментами при построениях не вытекает из существа самой геометрии. Как уже давным-давно установили греческие математики, некоторые проблемы — скажем, удвоение куба — могут быть решены, например, с привлечением угольника (с прямым углом); можно изобрести всякие другие инструменты, помимо циркуля, которые позволили бы чертить эллипсы, гиперболы и более сложные кривые: тем самым область фигур, допускающих построение, была бы значительно расширена. Однако мы будем придерживаться прочно установившегося понимания выполнимости геометрических построений, подразумевая, что разрешено пользоваться только циркулем и линейкой.

ЧАСТЬ 1

Доказательства невозможности и алгебра

§ 1. Основные геометрические построения

1. Построение полей и извлечение квадратных корней. В порядке развития общих идей мы начнем с рассмотрения небольшого числа классических построений. Более углубленное изучение возможности геометрических построений неизбежно связано с переводом геометрической задачи на язык алгебры. Всякая проблема геометрического построения может быть схематизирована следующим образом: дано некоторое число отрезков, скажем, a, b, c, ...; требуется построить один или несколько отрезков x, y, ... Даже если на первый взгляд проблема имеет совсем иной вид, ее всегда можно переформулировать таким образом, чтобы она включилась в указанную схему. Искомые отрезки фигурируют или в виде сторон треугольника, который требуется построить, или в виде радиусов кругов, или как прямоугольные координаты каких-то искомых точек (см., например, стр. 145). Предположим для простоты, что требуется построить какой-то отрезок x. В таком случае геометрическое построение приводит к решению алгебраической задачи: установить соотношение (в форме уравнения) между искомой величиной x и данными величинами a, b, c, ...; затем, решая это уравнение, найти формулу для величины x и, наконец, выяснить, можно ли свести вычисление x

к таким алгебраическим процедурам, которые соответствуют построениям, выполнимым с помощью циркуля и линейки. Таким образом, в основе всей рассматриваемой теории лежит принцип аналитической геометрии — количественная характеристика геометрических объектов, основанная на введении континуума действительных чисел.

Заметим прежде всего, что простейшие алгебраические операции соответствуют элементарным геометрическим построениям. Если даны два отрезка, длины которых равны a и b (измерение производится посредством «единичного» отрезка), то очень легко построить $a+b$, $a-b$, ra (где r — рациональное число), $\frac{a}{b}$ и ab.

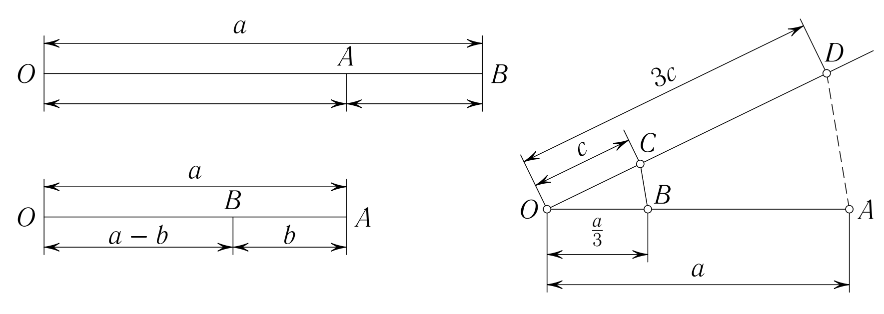


Рис. 27. Построение $a+b$ и $a-b$

Рис. 28. Построение $\frac{a}{3}$

Чтобы построить $a+b$ (рис. 27), мы проводим прямую линию и на ней откладываем циркулем отрезки $OA=a$ и $AB=b$. Тогда $OB=a+b$. Точно так же в случае $a-b$ мы откладываем $OA=a$ и $AB=b$, но на этот раз откладываем b в сторону, противоположную той, в которую отложили a. Тогда $OB=a-b$. Чтобы построить $3a$, мы просто строим $a+a+a$; аналогично поступаем, если нужно построить pa, где p — целое число. Отрезок $\frac{a}{3}$ строится следующим приемом (рис. 28): на произвольной прямой откладываем $OA=a$ и затем проводим другую прямую через точку O. На этой прямой откладываем произвольный отрезок $OC=c$ и строим $OD=3c$. Соединяем A и D прямой линией и проводим через точку C прямую, параллельную AD; пусть эта прямая пересекает OA в точке B. Треугольники OBC и OAD подобны; значит, $\frac{OB}{a}=\frac{OB}{OA}=\frac{OC}{OD}=\frac{1}{3}$ и $OB=\frac{a}{3}$. Таким же образом можно вообще построить $\frac{a}{q}$, где q — целое. Совершая эту операцию над отрезком pa, мы построим ra, где $r=\frac{p}{q}$ — какое угодно рациональное число.

Чтобы построить $\frac{a}{b}$ (рис. 29), откладываем $OB=b$ и $OA=a$ на сторонах произвольного угла с вершиной O и на стороне OB откладываем отрезок $OD=1$. Через D проводим прямую, параллельную AB; пусть

она пересекает OA в точке C. Тогда будем иметь: $OC = \frac{a}{b}$. Построение ab показано на рис. 30; здесь AD — прямая, проходящая через A и параллельная BC.

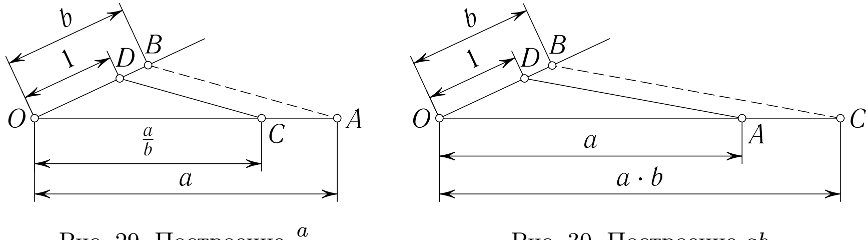


Рис. 29. Построение $\frac{a}{b}$ Рис. 30. Построение ab

Из этих соображений вытекает, что *«рациональные» алгебраические операции* — сложение, вычитание, умножение и деление, — производимые над заданными величинами, *могут быть выполнены посредством геометрических построений.* Исходя из данных отрезков, измеряемых действительными числами a, b, c, ..., мы можем, последовательно выполняя эти простые построения, построить любую величину, которая через a, b, c, ... выражается рационально, т. е. с помощью лишь перечисленных выше четырех основных действий. Совокупность всех величин, которые таким образом могут быть получены из a, b, c, ..., образует то, что называется *числовым полем* — множество чисел, обладающее тем свойством, что любая рациональная операция, совершенная над двумя (или более) элементами этого множества, приводит снова к элементу этого же множества. Мы напоминаем, что совокупность всех рациональных чисел, совокупность всех действительных чисел, совокупность всех комплексных чисел образуют такие поля. В рассматриваемом нами теперь случае говорят, что поле *порождается* данными числами a, b, c, ...

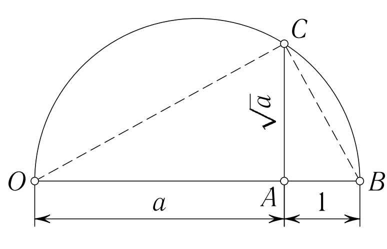


Рис. 31. Построение $\sqrt{a}$

Существенно новой операцией, выводящей нас за пределы полученного поля, является извлечение квадратного корня. Если задан отрезок a, то отрезок $\sqrt{a}$ может быть построен с помощью только циркуля и линейки. На произвольной прямой мы откладываем $OA = a$ и $AB = 1$ (рис. 31). Проводим, далее, окружность с диаметром OB и из точки A восставляем перпендикуляр к OB; пусть он пересекает окружность в точке C. Угол C в треугольнике OBC прямой (согласно теореме, известной из элементарной геометрии: угол, вписанный в полуокружность, прямой).

Значит, $\angle OCA = \angle ABC$, прямоугольные треугольники OAC и CAB подобны, и, полагая $AC = x$, мы получаем

$$\frac{a}{x} = \frac{x}{1}, \qquad x^2 = a, \qquad x = \sqrt{a}.$$

2. Правильные многоугольники. Рассмотрим теперь несколько более сложные конструктивные задачи. Начнем с построения *правильного десятиугольника.* Предположим, что правильный десятиугольник вписан в круг радиуса 1 (рис. 32); обозначим его сторону через x. Так как центральный угол, под которым эта сторона x видна из центра круга, содержит 36°, то остальные два угла большого треугольника содержат каждый по 72°, и значит, пунктирная линия, делящая пополам угол A, разбивает треугольник OAB на два равнобедренных треугольника с равными боковыми сторонами длины x. Радиус круга, таким образом, составляется из отрезков x и $1 - x$. Так как треугольник OAB подобен меньшему из двух треугольников, на которые он разбивается, то мы получаем $\dfrac{1}{x} = \dfrac{x}{1-x}$.

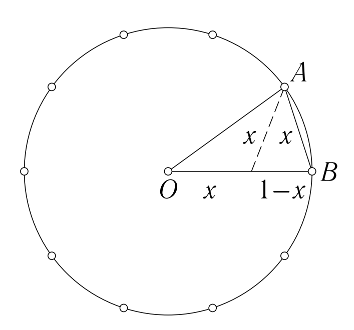


Рис. 32. Правильный десятиугольник

Эта пропорция приводит к квадратному уравнению $x^2 + x - 1 = 0$, решение которого имеет вид $x = \dfrac{\sqrt{5}-1}{2}$. (Другое решение нас не интересует, так как оно соответствует отрицательному значению x.) Из полученной формулы ясно, что отрезок x может быть построен геометрически. Имея же отрезок x, мы сможем построить правильный десятиугольник, откладывая по окружности десять раз хорду x. Отсюда уже легко получить и правильный пятиугольник, соединяя вершины десятиугольника через одну.

Вместо того чтобы строить $\sqrt{5}$ тем методом, который указан на рис. 31, мы можем построить гипотенузу прямоугольного треугольника со сторонами 1 и 2. Затем нужно отнять единичный отрезок и то, что получится, разделить пополам.

Отношение $\dfrac{OB}{AB}$ в рассмотренной задаче было названо «золотым», так как, по мнению греческих математиков, прямоугольник, стороны которого находятся в этом отношении, эстетически особенно приятен для глаза. Значение отношения приблизительно равно 1,62.

Из всех правильных многоугольников легче всего построить шестиугольник. Так как длина стороны такого шестиугольника, вписанного в круг, равна радиусу круга, то сам шестиугольник строится без за-

труднений, если мы отложим шесть раз по окружности отрезок, равный радиусу.

Имея правильный n-угольник, можно сейчас же получить и правильный $2n$-угольник, деля пополам дуги между соседними вершинами n-угольника. Начиная с диаметра круга (правильного вписанного «двуугольника»), мы построим последовательно 4, 8, 16, ..., 2^n-угольники. Таким же образом, начиная с шестиугольника, мы получим 12, 24, 48, ...-угольники, а начиная с десятиугольника, — 20, 40, ...-угольники.

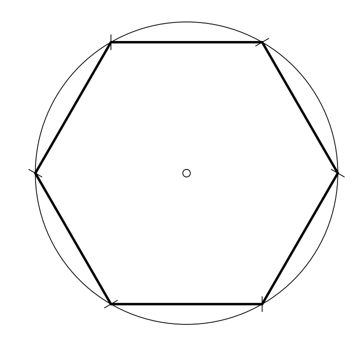


Рис. 33. Правильный шестиугольник

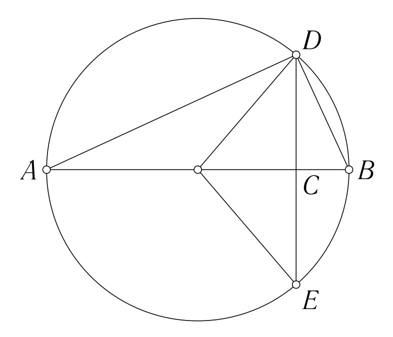


Рис. 34. Удвоение числа сторон правильного многоугольника

Если s_n обозначает длину стороны правильного n-угольника, вписанного в единичный круг (т. е. круг с радиусом 1), то сторона правильного вписанного $2n$-угольника будет иметь длину

$$s_{2n} = \sqrt{2 - \sqrt{4 - s_n^2}}.$$

Доказывается это следующим образом (рис. 34): пусть $s_n = DE = 2DC$, $s_{2n} = DB$ и $AB = 2$. Площадь прямоугольного треугольника ABD равна $\frac{1}{2}\,BD \cdot AD$, или, с другой стороны, $\frac{1}{2}\,AB \cdot CD$. Так как $AD = \sqrt{AB^2 - DB^2}$, то, подставляя $AB = 2$, $BD = s_{2n}$, $CD = \frac{1}{2}\,s_n$ и сравнивая между собой два выражения для площади, мы получаем

$$s_n = s_{2n}\sqrt{4 - s_{2n}^2}, \quad \text{или} \quad s_n^2 = s_{2n}^2(4 - s_{2n}^2).$$

Остается решить квадратное уравнение относительно $x = s_{2n}^2$ и при выборе корня принять во внимание, что x должно быть меньше 2.

Из этой формулы, так как длина s_4 (сторона квадрата) равна $\sqrt{2}$, следует, что

$$s_8 = \sqrt{2 - \sqrt{2}}, \qquad s_{16} = \sqrt{2 - \sqrt{2 + \sqrt{2}}},$$

$$s_{32} = \sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{2}}}} \quad \text{и т. д.}$$

В качестве общей формулы мы получаем (при $n > 2$)

$$s_{2^n} = \sqrt{2 - \sqrt{2 + \sqrt{2 + \ldots + \sqrt{2}}}},$$

причем в правой части должно быть всего $n - 1$ радикалов. Периметр 2^n-угольника, вписанного в круг радиуса 1, равен $2^n s_{2^n}$. Когда n стремится к бесконечности, этот периметр в пределе переходит в длину окружности, по определению равную 2π:

$$2^n s_{2^n} \to 2\pi \quad \text{при } n \to \infty.$$

Деля на два и подставляя m вместо $n - 1$, мы получаем следующую формулу для π:

$$2^m \underbrace{\sqrt{2 - \sqrt{2 + \sqrt{2 + \ldots + \sqrt{2}}}}}_{m \text{ радикалов}} \to \pi \quad \text{при } m \to \infty.$$

Упражнение. Пользуясь тем, что $2^m \to \infty$, докажите, как следствие, что

$$\underbrace{\sqrt{2 + \sqrt{2 + \ldots + \sqrt{2}}}}_{n \text{ радикалов}} \to 2 \quad \text{при } n \to \infty.$$

Резюмируем полученные здесь результаты таким образом: *стороны вписанных в единичный круг правильных 2^n-угольников, $5 \cdot 2^n$-угольников и $3 \cdot 2^n$-угольников вычисляются посредством рациональных операций — сложения, вычитания, умножения, деления — и операции извлечения квадратного корня; следовательно, они могут быть построены с помощью только циркуля и линейки.*

3. Проблема Аполлония. Другая конструктивная проблема, решающаяся весьма просто, если подойти к ней с алгебраической точки зрения, — это знаменитая и уже упомянутая выше проблема Аполлония о проведении окружности, касательной к трем данным окружностям. В настоящем контексте нам не представляется необходимым искать ее особенно элегантное решение. Нам существенно лишь установить принципиально важное положение: проблема Аполлония решается с помощью циркуля и линейки. Мы вкратце приведем соответствующее доказательство; вопрос же о наиболее элегантном построении будет разобран ниже (см. стр. 181).

Пусть центры трех данных окружностей имеют соответственно координаты (x_1, y_1), (x_2, y_2) и (x_3, y_3), а радиусы равны r_1, r_2 и r_3. Обозначим координаты центра искомой окружности через (x, y), а радиус через r. Легко написать условие касания двух окружностей, если учесть, что расстояние между центрами должно равняться сумме или разности радиусов, смотря по тому, имеет ли место внешнее или внутреннее касание. Записывая в алгебраической форме три условия задачи, мы получаем

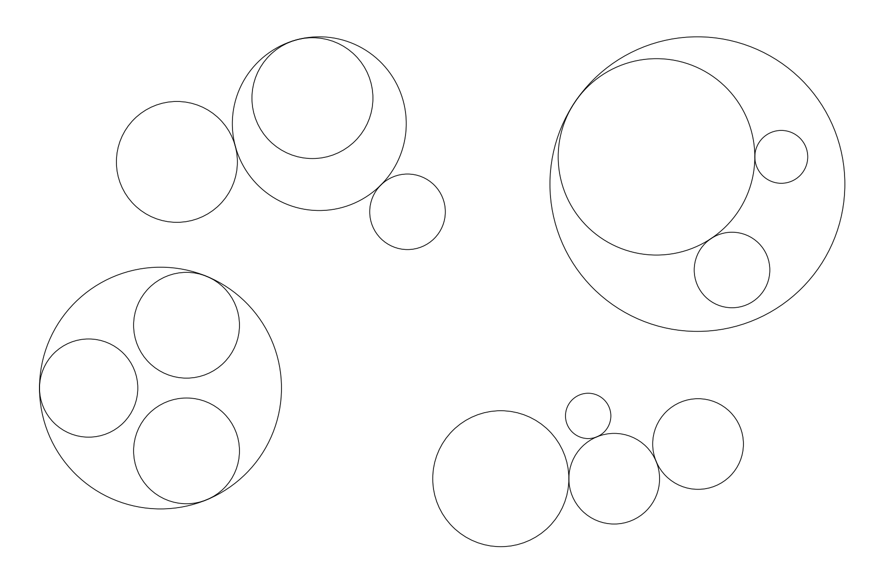


Рис. 35. Окружности Аполлония

три уравнения

$$(x - x_1)^2 + (y - y_1)^2 - (r \pm r_1)^2 = 0, \qquad (1)$$

$$(x - x_2)^2 + (y - y_2)^2 - (r \pm r_2)^2 = 0, \qquad (2)$$

$$(x - x_3)^2 + (y - y_3)^2 - (r \pm r_3)^2 = 0, \qquad (3)$$

которые после преобразований принимают вид

$$x^2 + y^2 - r^2 - 2xx_1 - 2yy_1 \pm 2rr_1 + x_1^2 + y_1^2 + r_1^2 = 0 \qquad (1а)$$

и т. п.

В каждом из уравнений нужно брать знак плюс или минус, в зависимости от того, каково касание — внешнее или внутреннее (рис. 35). Все уравнения (1), (2), (3) — второй степени относительно неизвестных x, y, r, но они обладают тем свойством, что члены второй степени входят в одинаковой комбинации, как видно из развернутой формы (1а). Таким образом, вычитая (2) из (1), мы получаем уравнение, линейное относительно x, y, r:

$$ax + by + cr = d, \qquad (4)$$

где $a = 2(x_2 - x_1)$ и т. д. Точно так же, вычитая (3) из (1), будем иметь другое линейное уравнение

$$a'x + b'y + c'r = d'. \qquad (5)$$

Решая уравнения (4) и (5) относительно неизвестных x и y, которые, таким образом, выразятся линейно через r, и затем подставляя в (1),

придем к уравнению, квадратному относительно r, каковое может быть решено с помощью рациональных операций и извлечения корня (см. стр. 143). Это уравнение, вообще говоря, будет иметь два решения, из которых лишь одно будет положительным. Определив r, найдем дальше значения x и y, подставляя r в ранее полученные формулы. Окружность с центром (x, y) и радиусом r должна быть касательной к трем данным окружностям. Во всей процедуре решения участвуют только рациональные операции и извлечение квадратного корня. Отсюда следует, что построение x, y и r может быть выполнено с помощью только циркуля и линейки.

В общем случае будет иметься 8 решений проблемы Аполлония в соответствии с возможными $2 \cdot 2 \cdot 2 = 8$ комбинациями в выборе знаков $+$ и $-$ в уравнениях (1), (2) и (3); выбор же знаков надлежит делать в зависимости от того, какого рода касание — внешнее или внутреннее — желательно иметь по отношению к каждой из данных окружностей. Вполне возможно, что наша алгебраическая процедура не приведет к действительным значениям x, y и r. Таков будет, например, случай, когда все три данные окружности — концентрические; тогда, очевидно, наша геометрическая задача не будет иметь ни одного решения. Следует также предвидеть возможность и случаев «вырождения»; например, если все три окружности «вырождаются» в точки, лежащие на одной прямой, тогда аполлониева окружность тоже «вырождается» в эту самую прямую. Мы не видим необходимости рассматривать вопрос во всех подробностях: это сделает сам читатель, если обладает некоторыми алгебраическими навыками.

§ 2. Числа, допускающие построение, и числовые поля

1. Общая теория. В предыдущем изложении мы постарались охарактеризовать общий, так сказать, алгебраический фон геометрических построений. Каждое геометрическое построение представляет ряд последовательных этапов из числа следующих: 1) проведение прямой линии через две точки, 2) нахождение точки пересечения двух прямых, 3) проведение окружности с данным центром и радиусом, 4) нахождение точки пересечения окружности с другой окружностью или прямой линией. Элемент (точка, прямая, окружность) считается известным в том случае, если он задается условием задачи или если он построен на предыдущей стадии задачи. Проводя теоретический анализ задачи, мы относим всю рассматриваемую конструкцию к некоторой координатной системе x, y (см. стр. 92). Тогда заданные элементы изображаются в виде точек или отрезков в плоскости x, y. Если задан только один отрезок, его можно принять в качестве единичного, в результате чего

фиксируется точка $x = 1$, $y = 0$. Иногда в процессе построения возникают произвольные элементы: проводятся произвольные прямые, строятся произвольные точки или круги. (Пример произвольного элемента мы имеем при нахождении середины отрезка: мы проводим два круга с центрами в концах отрезка и с одинаковыми, но произвольными радиусами, затем соединяем точки их пересечения.) В подобных случаях всегда можно считать произвольный элемент рациональным: произвольную точку можно выбрать так, чтобы у нее были рациональные координаты, произвольную прямую $ax + by + c = 0$ так, чтобы у нее были рациональные коэффициенты a, b, c, произвольный круг — так, чтобы рациональными были координаты центра и радиус. Мы условимся, что если в построении участвуют произвольные элементы, мы будем выбирать их рациональными: раз эти элементы в самом деле произвольны, такой выбор не повлияет на результат построения.

Ради простоты допустим в ближайшем рассуждении, что в условии задачи задается только один элемент — отрезок длины 1. Тогда в соответствии с результатами § 1 мы можем построить с помощью циркуля и линейки все числа, получающиеся из единицы посредством рациональных операций, т. е. рациональные числа $\frac{r}{s}$, где r и s — целые числа. Система рациональных чисел «замкнута» по отношению к рациональным операциям: сумма, разность, произведение, частное (исключая, как всегда, деление на 0) двух рациональных чисел снова являются рациональными числами. Всякое множество чисел, обладающее таким свойством замкнутости по отношению к четырем рациональным операциям, мы назвали *числовым полем* (стр. 75).

Упражнение. Покажите, что каждое числовое поле во всяком случае содержит все рациональные числа. (*Указание*: если a есть какое-нибудь не равное нулю число из поля F, то $\frac{a}{a} = 1$ также принадлежит к F, а из 1 можно получить все рациональные числа посредством рациональных операций.)

Отправляясь от единицы, можно построить все рациональное числовое поле и, следовательно, все рациональные точки (т. е. точки, у которых обе координаты рациональны) в плоскости x, y. Дальше, с помощью циркуля можно построить новые, иррациональные числа вроде числа $\sqrt{2}$, которое, как мы знаем из главы II, § 2, находится уже за пределами рационального поля. Но построив $\sqrt{2}$, можно еще дальше с помощью «рациональных» построений (§ 1) получить все числа вида

$$a + b\sqrt{2}, \tag{1}$$

где a и b рациональные и, следовательно, сами допускают построение. Можно также построить и числа вида

$$\frac{a + b\sqrt{2}}{c + d\sqrt{2}} \quad \text{или} \quad (a + b\sqrt{2})(c + d\sqrt{2}),$$

где a, b, c, d — рациональные. Однако эти числа всегда можно написать в форме (1). В самом деле,

$$\frac{a+b\sqrt{2}}{c+d\sqrt{2}} = \frac{a+b\sqrt{2}}{c+d\sqrt{2}} \cdot \frac{c-d\sqrt{2}}{c-d\sqrt{2}} = \frac{ac-2bd}{c^2-2d^2} + \frac{bc-ad}{c^2-2d^2}\sqrt{2} = p+q\sqrt{2},$$

где p и q рациональные. (Знаменатель c^2-2d^2 отличен от нуля, так как из $c^2-2d^2=0$ следовало бы $\sqrt{2}=\frac{c}{d}$, что противоречит факту иррациональности $\sqrt{2}$.) Точно так же

$$(a+b\sqrt{2})(c+d\sqrt{2}) = (ac+2bd)+(bc+ad)\sqrt{2} = r+s\sqrt{2},$$

где r и s рациональные. Итак, все, что мы можем построить исходя из $\sqrt{2}$, это числа вида (1), где a и b — произвольные рациональные числа.

Упражнение. Напишите в форме (1) числа

$$\frac{p}{q}, \quad p+p^2, \quad (p-p^2)\frac{q}{r}, \quad \frac{pqr}{1-r^2}, \quad \frac{p-qr}{q+pr^2},$$

где положено

$$p=1+\sqrt{2}, \quad q=2-\sqrt{2}, \quad r=-3+\sqrt{2}.$$

Как показывает предшествующее рассуждение, числа (1) снова образуют поле. Это поле обширнее, чем поле рациональных чисел, и включает его как часть («подполе»). Но, конечно, новое поле менее обширно, чем поле всех действительных чисел. Обозначим через F_0 поле рациональных чисел, а через F_1 — поле чисел вида (1). Мы установили возможность построения каждого числа из «расширенного» поля F_1. Можно и дальше расширять область чисел, допускающих построение, например, таким образом: выберем число из поля F_1, скажем $k=1+\sqrt{2}$, и, извлекая из него корень, получим новое допускающее построение число

$$\sqrt{1+\sqrt{2}} = \sqrt{k}.$$

Это число, в свою очередь, порождает (§ 1) поле, состоящее из всех чисел вида

$$p+q\sqrt{k}, \tag{2}$$

где p и q теперь уже числа из поля F_1, т. е. вида $a+b\sqrt{2}$, где a, b из F_0, т. е. рациональные.

Упражнение. Представьте числа

$$(\sqrt{k})^3, \quad \frac{1+(\sqrt{k})^3}{1+\sqrt{k}}, \quad \frac{\sqrt{2}\sqrt{k}+\frac{1}{\sqrt{2}}}{(\sqrt{k})^3-3}, \quad \frac{(1+\sqrt{k})\cdot(2-\sqrt{k})\left(\sqrt{2}+\frac{1}{\sqrt{k}}\right)}{1+\sqrt{3}k}$$

в форме (2).

Все эти числа были построены в предположении, что первоначально был задан только один отрезок. Если задано два отрезка, то один из них можно принять за единичный. Предположим, что второй отрезок выражается через

первый в виде числа α. Тогда можно построить поле G, состоящее из всех чисел вида
$$\frac{a_m\alpha^m + a_{m-1}\alpha^{m-1} + \ldots + a_1\alpha + a_0}{b_n\alpha^n + b_{n-1}\alpha^{n-1} + \ldots + b_1\alpha + b_0},$$
где $a_0, \ldots, a_m$ и $b_0, \ldots, b_n$ — рациональные, а m и n — произвольные целые положительные числа.

Упражнение. Считая заданными отрезки 1 и α, выполните построения для
$$1 + \alpha + \alpha^2, \quad \frac{1+\alpha}{1-\alpha}, \quad \alpha^3.$$

Будем исходить теперь из более общего предположения, что мы умеем строить все числа некоторого числового поля F. Убедимся, что *применение одной линейки не выведет нас за пределы поля* F. Уравнение прямой, проходящей через две точки с координатами a_1, b_1 и a_2, b_2 из поля F, имеет вид $(b_1 - b_2)x + (a_2 - a_1)y + (a_1b_2 - a_2b_1) = 0$ (см. стр. 515 и далее); коэффициенты в этом уравнении рационально зависят от чисел из поля F и, следовательно, сами принадлежат полю F. Далее, если у нас имеются две прямые $\alpha x + \beta y + \gamma = 0$ и $\alpha' x + \beta' y + \gamma' = 0$ с коэффициентами из F, то координаты точки пересечения, получающиеся при решении системы этих уравнений, суть
$$x = \frac{\gamma\beta' - \beta\gamma'}{\alpha\beta' - \beta\alpha'}, \qquad y = \frac{\alpha\gamma' - \gamma\alpha'}{\alpha\beta' - \beta\alpha'}.$$
Так как и они тоже являются числами из F, то ясно, что применение одной только линейки не выведет нас за пределы F.

Упражнение. Прямые $x + \sqrt{2}y - 1 = 0$, $2x - y + \sqrt{2} = 0$ имеют коэффициенты, принадлежащие полю (1). Вычислите коэффициенты точки их пересечения и проверьте, что они также вида (1); соедините точки $(1, \sqrt{2})$ и $(\sqrt{2}, 1 - \sqrt{2})$ прямой линией $ax + by + c = 0$ и проверьте, что коэффициенты a, b, c имеют вид (1). То же сделайте по отношению к полю (2) для прямых
$$\sqrt{1+\sqrt{2}}x + \sqrt{2}y = 1, \quad (1+\sqrt{2})x - y = 1 - \sqrt{1+\sqrt{2}}$$
и для точек
$$(\sqrt{2}, -1), \quad (1+\sqrt{2}, \sqrt{1+\sqrt{2}}).$$

Но с помощью циркуля можно выбраться за пределы поля F. Для этой цели выберем в поле F такое число k, что число $\sqrt{k}$ уже не будет принадлежать F. Число $\sqrt{k}$ можно построить с помощью циркуля, так же как и все числа вида
$$a + b\sqrt{k}, \tag{3}$$
где a, b — произвольные числа из F. Сумма и разность двух таких чисел
$$a + b\sqrt{k} \quad \text{и} \quad c + d\sqrt{k},$$
их произведение
$$(a + b\sqrt{k})(c + d\sqrt{k}) = (ac + kbd) + (ad - bc)\sqrt{k}$$

и их отношение

$$\frac{a+b\sqrt{k}}{c+d\sqrt{k}} = \frac{(a+b\sqrt{k})(c-d\sqrt{k})}{c^2-kd^2} = \frac{ac-kbd}{c^2-kd^2} + \frac{bc-ad}{c^2-kd^2}\sqrt{k}$$

— снова числа вида $p+q\sqrt{k}$, где p и q принадлежат F. (Знаменатель $c^2 - kd^2$ не обращается в нуль, так как c и d одновременно не обращаются в нуль: иначе мы получили бы $\sqrt{k} = \frac{c}{d}$, что противоречит допущению, что $\sqrt{k}$ не принадлежит F.) Итак, множество чисел вида $a+b\sqrt{k}$ образует некоторое поле F'. Поле F' включает поле F как «подполе» (достаточно положить $b=0$). Будем называть F' «расширенным» полем.

В качестве примера рассмотрим поле F чисел вида $a+b\sqrt{2}$, где a, b рациональные: возьмем $k=\sqrt{2}$. Тогда числа расширенного поля F' имеют вид $p+q\sqrt[4]{2}$, где p и q принадлежат F, $p=a+b\sqrt{2}$, $q=a'+b'\sqrt{2}$, а числа a, b, a', b' — рациональные. Всякое число из F' может быть записано в этой форме, например,

$$\frac{1}{\sqrt{2}+\sqrt[4]{2}} = \frac{\sqrt{2}-\sqrt[4]{2}}{(\sqrt{2}+\sqrt[4]{2})(\sqrt{2}-\sqrt[4]{2})} = \frac{\sqrt{2}-\sqrt[4]{2}}{2-\sqrt{2}} = \frac{\sqrt{2}}{2-\sqrt{2}} - \frac{\sqrt[4]{2}}{2-\sqrt{2}} =$$
$$= \frac{\sqrt{2}(2+\sqrt{2})}{4-2} - \frac{(2+\sqrt{2})}{4-2}\sqrt[4]{2} = (1+\sqrt{2}) - \left(1+\frac{1}{2}\sqrt{2}\right)\sqrt[4]{2}.$$

Упражнение. Пусть F есть поле $p+q\sqrt{2+\sqrt{2}}$, где p, q — вида $a+b\sqrt{2}$, а числа a, b рациональные. Представьте $\dfrac{1+\sqrt{2+\sqrt{2}}}{2-3\sqrt{2+\sqrt{2}}}$ в таком же виде.

Мы убедились, что, отправляясь от некоторого поля F чисел, допускающих построение, и выбрав произвольное число k из этого поля, мы можем с помощью циркуля и линейки построить число $\sqrt{k}$, а значит, и все числа вида $a+b\sqrt{k}$, где a, b принадлежат F.

Покажем теперь, обратно, что, пользуясь только циркулем, мы можем получить числа *только* указанного вида. В самом деле, в результате однократного применения циркуля можно сделать только одно из двух: или найти точку пересечения окружности и прямой, или найти точку пересечения двух окружностей (то и другое равносильно построению координат точки пересечения). Окружность с центром (ξ, η) и радиусом r имеет уравнение $(x-\xi)^2 + (y-\eta)^2 = r^2$; поэтому, если ξ, η, r принадлежат F, то уравнение окружности, записанное в виде

$$x^2+y^2+2\alpha x+2\beta y+\gamma = 0,$$

будет иметь коэффициенты α, β, γ, принадлежащие также F. Прямая линия

$$ax+by+c=0,$$

соединяющая две точки с координатами F, имеет также коэффициенты из F (см. стр. 150). Исключая y из этих двух уравнений, мы получаем

для координаты x точки пересечения окружности и прямой квадратное уравнение вида
$$Ax^2 + By + C = 0$$
с коэффициентами A, B, C из F (именно, $A = a^2 + b^2$, $B = 2(ac + b^2\alpha - ab\beta)$, $C = c^2 - 2bc\beta + b^2\gamma$). Решение дается формулой
$$x = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A},$$
которая имеет вид $p + q\sqrt{k}$, где p, q, k принадлежат F. Такая же формула получается и для координаты y точки пересечения.

С другой стороны, если речь идет о двух окружностях
$$x^2 + y^2 + 2\alpha x + 2\beta y + \gamma = 0,$$
$$x^2 + y^2 + 2\alpha' x + 2\beta' y + \gamma' = 0,$$
то, вычитая одно уравнение из другого, мы получим линейное уравнение
$$(\alpha - \alpha')x + (\beta - \beta')y + (\gamma - \gamma') = 0,$$
которое можно решить совместно с одним из уравнений двух окружностей.

В обоих случаях построение дает нам обе координаты одной или двух новых точек, и эти новые величины имеют вид $p + q\sqrt{k}$, причем p, q, k принадлежат F. В частности, $\sqrt{k}$ может сам оказаться принадлежащим F, например, если $k = 4$. Но, вообще говоря, этого не будет.

Упражнение. Рассмотрим окружность с центром в начале координат и радиусом $2\sqrt{2}$ и прямую, соединяющую точки $\left(\frac{1}{2}, 0\right)$, $(4\sqrt{2}, \sqrt{2})$. Определите поле F', порождаемое точками пересечения окружности и прямой. Сделайте то же по отношению к точкам пересечения данной окружности с окружностью, у которой радиус равен $\frac{\sqrt{2}}{2}$, а центр есть $(0, 2\sqrt{2})$.

Подведем еще раз итоги. Отправляясь от некоторых заданных величин (отрезков или чисел), с помощью одной только линейки мы можем построить все величины из поля F, порождаемого данными величинами с помощью рациональных операций, но не выйдем за пределы этого поля. Воспользовавшись циркулем, мы расширяем поле величин, допускающих построение, и получаем новое расширенное поле F', состоящее из чисел вида $a + b\sqrt{k}$, где a, b, k принадлежат F. Поле F есть подполе поля F': всякое число из F принадлежит также F', так как в формуле $a + b\sqrt{k}$ можно положить $b = 0$. (Предполагается, что $\sqrt{k}$ есть новое число, не принадлежащее F; иначе F' совпало бы с F.) Мы убедились, что в результате каждого геометрического построения (т. е. проведения прямой через две известные точки; проведения окружности, имеющей известный центр и известный радиус; нахождения пересечения двух известных прямых или окружностей) или получаются величины,

принадлежащие первоначальному полю, или же, при построении квадратного корня, открывается новое, расширенное поле величин, допускающих построение.

Мы теперь в состоянии точно охарактеризовать совокупность всех величин, допускающих построение с помощью только циркуля и линейки. Будем исходить из некоторого поля F_0, определяемого величинами, входящими в условие задачи; например, это будет поле рациональных чисел, если задан только один отрезок, выбираемый в качестве единичного. Далее, «присоединяя» к полю величину $\sqrt{k_0}$, (где k_0 принадлежит F_0, но $\sqrt{k_0}$ ему не принадлежит), строим новое поле F_1 чисел, допускающих построение вида $a_0 + b_0\sqrt{k_0}$, где a_0, b_0 принадлежат F_0. Еще дальше, посредством «присоединения» $\sqrt{k_1}$ (где k_1 принадлежит F_1, но $\sqrt{k_1}$ не принадлежит), получается новое поле F_2 чисел вида $a_1 + b_1\sqrt{k_1}$, где a_1 и b_1 принадлежат F_1. Повторяя эту процедуру, приходим вообще к полю F_n после «присоединения» n квадратных корней. *С помощью только циркуля и линейки допускают построение те и только те числа, которые после конечного числа «присоединений» описанного выше типа включаются в расширенное поле F_n.* Число n необходимых «присоединений» не имеет особенно большого значения; но оно до некоторой степени характеризует, насколько сложна рассматриваемая проблема.

Иллюстрируем описанную процедуру следующим примером. Нужно построить число

$$\sqrt{6} + \sqrt{\sqrt{\sqrt{1+\sqrt{2}} + \sqrt{3}} + 5}.$$

Пусть F_0 — поле рациональных чисел. Полагая $k_0 = 2$, получаем поле F_1, содержащее число $1 + \sqrt{2}$. Возьмем затем $k_1 = 1 + \sqrt{2}$ и $k_2 = 3$. Число 3 содержится уже в начальном поле F_0, значит, и подавно в поле F_2, так что положить $k_2 = 3$ вполне допустимо. Потом возьмем $k_3 = \sqrt{1+\sqrt{2}} + \sqrt{3}$ и, наконец, $k_4 = \sqrt{\sqrt{1+\sqrt{2}} + \sqrt{3}} + 5$. Полученное после этого поле F_5 уже содержит интересующее нас число, так как $\sqrt{6}$ в нем содержится: действительно, $\sqrt{2}$ и $\sqrt{3}$, а следовательно, и их произведение, содержатся уже в F_3, значит, и подавно — в F_5.

Упражнение. Отправляясь от рационального поля, проверьте, что сторона правильного 2^m-угольника (см. стр. 145) допускает построение ($n = m - 1$). Проследите за тем, какова последовательность постепенно расширяемых полей. Сделайте то же самое с числами

$$\sqrt{1+\sqrt{2}+\sqrt{3}+\sqrt{5}}, \qquad \frac{\sqrt{5}+\sqrt{11}}{1+\sqrt{7-\sqrt{3}}},$$

$$\left(\sqrt{2+\sqrt{3}}\right)\left(\sqrt[8]{2} + \sqrt{1+\sqrt{2+\sqrt{5}}+\sqrt{3-\sqrt{7}}}\right).$$

2. Все числа, допускающие построение — алгебраические. Если начальное поле F_0 есть рациональное поле (порождаемое единственным отрезком), то все числа, допускающие построение, принадлежат к числу алгебраических. (Определение алгебраических чисел было дано на стр. 123.) Именно, числа поля F_1 являются корнями квадратных уравнений, числа поля F_2 — корнями уравнений четвертой степени, и вообще, числа поля F_k — корнями уравнений степени $2k$ с рациональными коэффициентами. Докажем это сначала для поля F_2, причем начнем с примера. Пусть $x = \sqrt{2} + \sqrt{3 + \sqrt{2}}$. Мы получаем $(x - \sqrt{2})^2 = 3 + \sqrt{2}$, $x^2 + 2 - 2\sqrt{2}x = 3 + \sqrt{2}$, или $x^2 - 1 = \sqrt{2}(2x + 1)$ — квадратное уравнение с коэффициентами из F_1. Возведение в квадрат приводит к уравнению

$$(x^2 - 1)^2 = 2(2x + 1)^2$$

четвертой степени с рациональными коэффициентами.

В общем случае любое число поля F_2 имеет вид

$$x = p + q\sqrt{w}, \tag{4}$$

где p, q, w принадлежат полю F_1 и, значит, имеют вид $p = a + b\sqrt{s}$, $q = c + d\sqrt{s}$, $w = e + f\sqrt{s}$, где a, b, c, d, e, f, s — рациональные числа. Из равенства (4) мы получаем

$$x^2 - 2px + p^2 = q^2 w,$$

причем все коэффициенты принадлежат полю F_1, порождаемому величиной $\sqrt{s}$. Поэтому последнее равенство можно переписать в виде

$$x^2 + ux + v = \sqrt{s}(rx + t),$$

где коэффициенты r, s, t, u, v — рациональные. Возводя в квадрат, получим уравнение четвертой степени

$$(x^2 + ux + v)^2 = s(rx + t)^2 \tag{5}$$

с рациональными коэффициентами, как и требовалось.

Упражнения. 1) Постройте уравнения с рациональными коэффициентами для чисел

а) $x = \sqrt{2 + \sqrt{3}}$, б) $x = \sqrt{2} + \sqrt{3}$, в) $x = \dfrac{1}{\sqrt{5 + \sqrt{3}}}$.

2) Постройте таким же образом уравнения восьмой степени для чисел

а) $x = \sqrt{2 + \sqrt{2 + \sqrt{2}}}$, б) $x = \sqrt{2} + \sqrt{1 + \sqrt{3}}$,

в) $x = 1 + \sqrt{5 + \sqrt{3 + \sqrt{2}}}$.

Чтобы закончить доказательство теоремы в общем случае, когда x принадлежит полю F_k с произвольным индексом k, достаточно установить, как выше, что x удовлетворяет квадратному уравнению с коэффициентами из поля F_{k-1}. Затем, повторяя процедуру доказательства, убеждаемся, что x удовлетворяет уравнению степени $2^2 = 4$ с коэффициентами из поля F_{k-2}, и т. д.

Упражнение. Закончите это общее доказательство, применяя метод математической индукции: докажите, что x удовлетворяет уравнению степени 2^l с коэффициентами из поля F_{k-l}, $0 < l \leqslant k$. При $l = k$ получается окончательный результат.

§ 3. Неразрешимость трех классических проблем

1. Удвоение куба. Теперь мы уже достаточно подготовлены к исследованию известных еще с древности проблем трисекции угла, удвоения куба и построения правильного семиугольника. Рассмотрим прежде всего проблему удвоения куба.

Если данный куб имеет ребро, равное единице, его объем будет равен кубической единице; требуется найти ребро x куба, объем которого вдвое больше. Итак, искомое ребро удовлетворяет простому кубическому уравнению

$$x^3 - 2 = 0. \tag{1}$$

Наше доказательство невозможности построения числа x с помощью только циркуля и линейки будет носить «косвенный» характер. Допустим, что такое построение возможно. Тогда, согласно полученным выше результатам, число x должно принадлежать некоторому полю F_k, полученному так, как было объяснено раньше, — из рационального поля посредством последовательного «присоединения» квадратных корней. Мы сейчас убедимся в том, что такое допущение приведет к противоречию.

Мы уже знаем, что число x не может принадлежать рациональному полю F_0, так как $\sqrt[3]{2}$ есть число иррациональное (см. упражнение 1 на стр. 80). Значит, придется допустить, что оно принадлежит одному из расширенных полей F_k, где k — целое положительное число. Мы имеем право допустить, что k есть наименьшее из таких целых чисел, т. е. что x принадлежит F_k, но не принадлежат F_{k-1}. Это значит, что x имеет вид

$$x = p + q\sqrt{w},$$

где p, q и w принадлежат какому-то полю F_{k-1}, но $\sqrt{w}$ ему не принадлежит. Основываясь, далее, на довольно простом алгебраическом рассуждении (подобные рассуждения приходится применять нередко), мы убедимся, что если $p + q\sqrt{w}$ есть решение уравнения (1), то $y = p - q\sqrt{w}$ есть также его решение. Так как x принадлежит полю F_k, то x^3 и $x^3 - 2$ тоже принадлежат F_k и, значит,

$$x^3 - 2 = a + b\sqrt{w}, \tag{2}$$

где a и b принадлежат F_{k-1}. Нетрудно подсчитать, что $a = p^3 + 3pq^2w - 2$, $b = 3p^2q + q^3w$. Если положим

$$y = p - q\sqrt{w},$$

то сразу видно, что

$$y^3 - 2 = a - b\sqrt{w}. \tag{2'}$$

Так как мы предположили, что x есть корень уравнения (1), то

$$a + b\sqrt{w} = 0. \tag{3}$$

Но из последнего равенства следует (это основной момент рассуждения!), что оба числа a и b равны нулю. Действительно, если бы b было отлично от нуля, то из (3) получилось бы равенство $\sqrt{w} = -\frac{a}{b}$, это противоречит допущению, что $\sqrt{w}$ не принадлежит полю F_{k-1}. Итак, $b = 0$, и тогда из (3) следует, что $a = 0$. Но раз мы установили, что $a = b = 0$, то уже из равенства (2′) немедленно вытекает, что $y = p - q\sqrt{w}$ есть решение уравнения (1), так как $y^3 - 2 = 0$. Далее $y \neq x$, т. е. $x - y \neq 0$, так как число $x - y = 2q\sqrt{w}$ могло бы обращаться в нуль только при $q = 0$, а в этом случае $x = p$ принадлежало бы полю F_{k-1}, чего мы не предполагали.

Мы установили, что если $x = p + q\sqrt{w}$ есть корень кубического уравнения (1), то $y = p - q\sqrt{w}$ есть другой, не равный ему, корень того же уравнения. Но это немедленно приводят к противоречию: $y = p - q\sqrt{w}$ есть, очевидно, действительное число, так как числа p, q, $\sqrt{w}$ действительные, уравнение же (1) имеет только один действительный корень, а два — мнимых (см. стр. 119).

Наше первоначальное допущение привело к противоречию, значит, оно ошибочно; поэтому корень уравнения (1) не может принадлежать никакому полю F_k. Итак, удвоение куба с помощью только циркуля и линейки невозможно.

2. Одна теорема о кубических уравнениях. Заключительная часть только что приведенного алгебраического рассуждения была приспособлена к специальному уравнению, которым мы занимались. Но если мы хотим исследовать две другие проблемы древности, то желательно основываться на некоторой теореме общего характера. С алгебраической точки зрения все три проблемы связаны с решением кубического уравнения. Отлично известно, что если x_1, x_2, x_3 — три корня кубического уравнения

$$z^3 + az^2 + bz + c = 0, \tag{4}$$

то они связаны между собой соотношением

$$x_1 + x_2 + x_3 = -a. \tag{5}$$ [1]

Рассмотрим кубическое уравнение (4), в котором коэффициенты a, b, c пусть будут рациональными числами. Может, конечно, случиться, что

---

[1] Многочлен $z^3 + az^2 + bz + c$ можно представить в виде произведения трех множителей $(z - x_1)(z - x_2)(z - x_3)$, где x_1, x_2, x_3 — корни уравнения (4) (см. стр. 122). Отсюда следует тождество

$$z^3 + az^2 + bz + c = z^3 - (x_1 + x_2 + x_3)z^2 + (x_1x_2 + x_2x_3 + x_1x_3)z - x_1x_2x_3,$$

и так как коэффициенты при одинаковых степенях должны быть равны между собой, то

$$-a = x_1 + x_2 + x_3, \quad b = x_1x_2 + x_2x_3 + x_1x_3, \quad c = -x_1x_2x_3.$$ — *Прим. ред.*

один из корней уравнения есть рациональное число: например, уравнение $x^3 - 1 = 0$ имеет один корень 1 — рациональный, тогда как два других, удовлетворяющих квадратному уравнению $x^2 + x + 1 = 0$, — мнимые. Но мы сейчас докажем такую общую теорему: *если кубическое уравнение с рациональными коэффициентами не имеет рациональных корней, то ни один из его корней не может быть построен с помощью циркуля и линейки, исходя из рационального поля* F_0.

Доказательство будем вести, как раньше, косвенным методом. Допустим, что число x, являющееся корнем уравнения (4), допускает построение. Тогда x должно принадлежать некоторому полю F_k, последнему в цепи постепенно расширяемых полей $F_0, F_1, \ldots, F_k$.

Мы, как раньше, имеем право допустить, что никакой корень уравнения (4) не принадлежит полю F_{k-1}. (Что k не есть нуль, следует как раз из условия теоремы: x не может быть рациональным числом.) Итак, x может быть записано в виде

$$x = p + q\sqrt{w},$$

причем p, q, w принадлежат полю F_{k-1}, но $\sqrt{w}$ не принадлежит F_{k-1}. Такое же самое рассуждение, какое было проведено в предыдущем пункте, приводит к заключению, что число

$$y = p - q\sqrt{w},$$

также принадлежащее F_k, является корнем уравнения (4). Мы видим, как раньше, что $q \neq 0$; значит, $x \neq y$.

Из равенства (5) мы теперь заключаем, что третий корень уравнения (4) дается формулой $u = -a - x - y$. Но так как $x + y = 2p$, то, значит,

$$u = -a - 2p.$$

Радикал $\sqrt{w}$ здесь исчез, так что оказывается, что u принадлежит полю F_{k-1}. Это противоречит сделанному допущению, согласно которому k есть наименьшее целое число такое, что некоторое поле F_k содержит корень уравнения (4). Придется отвергнуть сделанное допущение, раз оно привело к противоречию, и признать, что ни один из корней уравнения (4) не принадлежит никакому полю F_k. Теорема доказана. На основании этой теоремы можно утверждать, что некоторое число не может быть построено с помощью только циркуля и линейки, как только установлено, что это число является корнем кубического уравнения с рациональными коэффициентами, не имеющего рациональных корней.

Теперь мы можем перейти к рассмотрению двух других проблем древности; заметим, что каждая из них облекается в алгебраическую форму не столь непосредственно, как уже рассмотренная.

3. Трисекция угла. Покажем, что трисекция угла с помощью только циркуля и линейки в *общем случае* невозможна. Конечно, существуют углы, например углы в 90° или в 180°, для которых трисекция выполняется. Но мы должны показать, что не существует процедуры построения, пригодной для *всякого* угла. Так как *общий* метод должен был бы относиться ко *всем* углам, то наша цель будет достигнута, если мы укажем хотя бы один какой-нибудь угол, для которого трисекция невозможна. Итак, несуществование общего метода трисекции будет установлено, если мы убедимся, что, например, угол в 60° не может быть разделен на три равные части с помощью только циркуля и линейки.

Алгебраический эквивалент рассматриваемой проблемы можно получить разными способами; самый простой способ — считать, что угол θ задан своим косинусом: $\cos\theta = g$. Тогда проблема сводится к вычислению величины $x = \cos\frac{\theta}{3}$. Интересующие нас косинусы связаны между собой простой тригонометрической формулой (см. стр. 118)

$$\cos\theta = 4\cos^3\frac{\theta}{3} - 3\cos\frac{\theta}{3}.$$

Другими словами, проблема трисекции угла θ (такого, что $\cos\theta = g$) равносильна построению корня кубического уравнения

$$4z^3 - 3z - g = 0. \tag{6}$$

Так как по предыдущему мы имеем право положить $\theta = 60^\circ$, $g = \cos 60^\circ = \frac{1}{2}$, то уравнение (6) принимает вид

$$8z^3 - 6z = 1. \tag{7}$$

В силу теоремы, доказанной в предыдущем пункте, для нашей цели достаточно показать, что это уравнение не имеет рациональных корней. Положим $v = 2z$; уравнение примет еще более простой вид

$$v^3 - 3v = 1. \tag{8}$$

Если бы существовало рациональное число $v = \frac{r}{s}$, удовлетворяющее этому уравнению, где r и s — целые числа без общего множителя (> 1), то мы должны были бы иметь равенство $r^3 - 3s^2r = s^3$. Отсюда следовало бы, что число $s^3 - r(r^2 - 3s^2)$ делится на r, и тогда получилось бы, что r и s имеют общий множитель, если только r не равно ± 1. Совершенно так же мы заключили бы, что число $r^3 = s^2(s + 3r)$ делится на s^2, а это значило бы, что r и s имеют общий множитель, если только s не равно ± 1. Но так как дробь $\frac{r}{s}$ по предположению несократима, то, значит, остается заключить, что числа r и s равны ± 1, т. е. $v = \pm 1$. Но подставляя $v = +1$ и $v = -1$ в уравнение (8), мы видим, что в обоих случаях уравнение не

удовлетворяется. Итак, уравнение (8), а следовательно, и уравнение (7) не имеют рациональных корней; тем самым невозможность трисекции угла доказана.

Эта теорема доказана в предположении, что линейка рассматривается как инструмент, служащий для проведения прямой через две данные точки, *и никак иначе.* В самом деле, когда мы давали общую характеристику чисел, которые допускают построение, имелось в виду только такое употребление линейки. Если допустить иные приемы пользования линейкой, то совокупность выполнимых построений чрезвычайно расширяется. Хорошим примером является следующий метод трисекции угла, указываемый в сочинениях Архимеда.

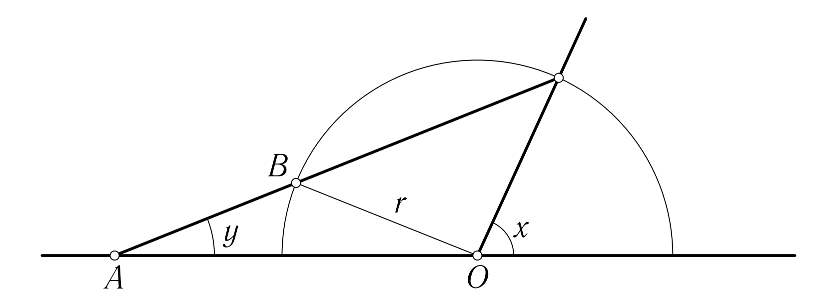


Рис. 36. Прием трисекции угла, указанный Архимедом

Пусть дан угол x (рис. 36). Продолжим горизонтальную сторону угла влево и затем проведем полукруг с центром O и произвольным радиусом r. Отметим *на самой линейке* такие точки A и B, что $AB = r$. Затем приведем линейку в такое положение, чтобы точка A линейки была на продолженной стороне угла, точка B на проведенном полукруге и вместе с тем линейка прошла бы через точку пересечения второй стороны угла с полукругом. В этом положении линейки проведем по ней прямую линию, образующую с продолженной стороной данного угла угол, который обозначим через y.

Упражнение. Докажите, что $y = \frac{x}{3}$.

4. Правильный семиугольник. Перейдем теперь к проблеме построения стороны x правильного семиугольника, вписанного в единичный круг. Проще всего справиться с этой проблемой, если прибегнуть к комплексным числам (см. главу II, § 5). Мы знаем, что вершины правильного семиугольника служат корнями уравнения

$$z^7 - 1 = 0, \tag{9}$$

причем координаты x, y каждой вершины являются действительной и мнимой частями комплексного числа $z = x + iy$. Один из корней есть $z = 1$, а остальные удовлетворяют уравнению

$$\frac{z^7 - 1}{z - 1} = z^6 + z^5 + z^4 + z^3 + z^2 + z + 1 = 0 \tag{10}$$

(см. стр. 120). Деля на z^3, получаем новое уравнение

$$z^3 + \frac{1}{z^2} + z^2 + \frac{1}{z^2} + z + \frac{1}{z} + 1 = 0. \tag{11}$$

Простые алгебраические преобразования приводят его к виду

$$\left(z + \frac{1}{z}\right)^3 - 3\left(z + \frac{1}{z}\right) + \left(z + \frac{1}{z}\right)^2 - 2 + \left(z + \frac{1}{z}\right) + 1 = 0. \tag{12}$$

Положив теперь

$$z + \frac{1}{z} = y,$$

мы приходим окончательно к уравнению третьей степени

$$y^3 + y^2 - 2y - 1 = 0. \tag{13}$$

Мы знаем, что z, корень седьмой степени из единицы, дается формулой

$$z = \cos\varphi + i\sin\varphi, \tag{14}$$

где $\varphi = \frac{360^\circ}{7}$ есть угол, под которым из центра круга видна сторона семиугольника; кроме того, из упражнения 2 на стр. 118 следует, что $\frac{1}{z} = \cos\varphi - i\sin\varphi$, так что

$$y = z + \frac{1}{z} = 2\cos\varphi.$$

Если мы сумеем построить y, то сумеем построить и $\cos\varphi$, и обратно. Итак, раз будет установлено, что величина y не может быть построена, то тем самым будет установлено, что не могут быть построены ни величина $\cos\varphi$, ни величина z; следовательно, невозможно будет построение семиугольника.

Таким образом, в силу теоремы пункта 2, остается показать, что уравнение (13) не имеет рациональных корней. Это тоже доказывается косвенным методом. Допустим, что уравнение (13) имеет рациональный корень $\frac{r}{s}$, где r и s — целые числа без общих множителей. В таком случае должно удовлетворяться равенство

$$r^3 + r^2 s - 2rs^2 - s^3 = 0; \tag{15}$$

отсюда ясно, что r^3 делится на s, а s^3 — на r. Так как r и s — взаимно простые числа, то отсюда следует, что каждое из них равно ± 1. Значит, и y, если только это число рациональное, должно равняться или $+1$ или -1. Но подстановка в уравнение (13) показывает, что ни $+1$, ни -1 не являются корнями уравнения. Итак, нельзя построить величины y, а следовательно, и стороны семиугольника.

5. Замечания по поводу квадратуры круга. Сравнительно элементарные методы позволили нам довести до конца исследование проблем удвоения куба, трисекции угла и построения правильного семиугольника. Но проблема квадратуры круга гораздо сложнее и требует

техники математического анализа. Так как круг радиуса r имеет площадь πr^2, то проблема построения квадрата, площадь которого равна площади круга с радиусом 1, равносильна построению числа $\sqrt{\pi}$, равного стороне искомого квадрата. Число $\sqrt{\pi}$ допускает построение в том и только том случае, если допускает построение число π. Исходя из данной нами общей характеристики чисел, допускающих построение, мы установили бы неразрешимость проблемы квадратуры круга, если бы показали, что π не содержится ни в каком поле F_k, возникающем из поля рациональных чисел посредством последовательных присоединений квадратных корней. Так как все числа, принадлежащие таким полям, являются алгебраическими, т. е. удовлетворяющими алгебраическим уравнениям с целыми коэффициентами, то неразрешимость квадратуры круга была бы доказана, если бы было установлено, что число π не алгебраическое, а трансцендентное (см. стр. 124).

Технический аппарат, необходимый для доказательства трансцендентности числа π, был создан Шарлем Эрмитом (1822–1905), который доказал вместе с тем трансцендентность числа e. Несколько усовершенствовав метод Эрмита, Ф. Линдеман (в 1882 г.) сумел доказать трансцендентность числа π и тем самым окончательно исчерпал вопрос, остававшийся без ответа на протяжении тысячелетий. Доказательство Линдемана — вне пределов, намеченных для этой книги, хотя оно и по плечу учащемуся, несколько знакомому с математическим анализом.

ЧАСТЬ 2

Различные методы выполнения построений

§ 4. Геометрические преобразования. Инверсия

1. Общие замечания. В настоящей, второй части этой главы мы систематически рассмотрим некоторые общие принципы, которые могут быть приложены к конструктивным проблемам. Многие из этих проблем обозреваются гораздо легче, если смотреть на них с общей точки зрения «геометрических преобразований». Вместо того чтобы изучать отдельное построение, мы займемся сразу целым классом проблем, связанных между собой теми или иными процедурами преобразований. Способность бросать яркий свет на существо вещей, присущая идее класса геометрических преобразований, никоим образом не ограничена конструктивными проблемами, но имеет ближайшее отношение ко всей геометрии в целом. В главах IV

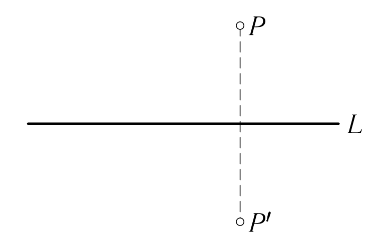


Рис. 37. Отражение точки относительно прямой

и V мы будем иметь случай оценить роль геометрических преобразований в этом более широком аспекте. Пока же мы подвергнем изучению один из частных типов преобразований — *инверсию плоскости относительно окружности*, представляющую собой обобщение обыкновенного зеркального отражения относительно прямой линии.

Говоря о *преобразовании* (*отображении*) плоскости самой в себя, мы имеем в виду некоторое правило, сопоставляющее каждой точке P плоскости некоторую другую точку P' той же плоскости. Точка P' называется *образом* точки P, точка P — *прообразом* точки P'. Простейший пример такого преобразования — *зеркальное отражение* (осевая симметрия) плоскости относительно данной прямой линии L: точка P по одну сторону L имеет своим образом точку P', расположенную по другую сторону L таким образом, что L является перпендикуляром к отрезку PP', восставленным из его середины. Преобразование может оставлять некоторые точки плоскости неподвижными; в нашем примере таковы точки самой прямой L.

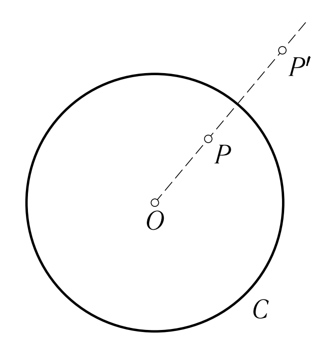


Рис. 38. Инверсия точки относительно окружности

Дальнейшими примерами преобразований являются *вращения* плоскости относительно неподвижной точки O, затем *параллельные переносы*, перемещающие каждую точку в данном направлении на одно и то же расстояние (это преобразование не имеет неподвижных точек), и, наконец, в качестве несколько более общего примера следует назвать *движения* плоскости, которые можно представлять себе составленными из вращений и параллельных переносов.

Но в данный момент нас интересует иной, частный класс преобразований, — именно, *инверсии* относительно окружностей. (Иногда их называют круговыми отражениями, вследствие наличия приблизительного сходства с отражением в сферическом зеркале.) Пусть в неподвижной плоскости задана некоторая окружность C с центром O (называемым *центром*, или *полюсом*, *инверсии*) и радиусом r. Образ точки P опре-

деляется как точка P', лежащая на прямой OP по ту же сторону от O, что и P, и такая, что

$$OP \cdot OP' = r^2. \qquad (1)$$

Из этого определения следует, что если P' есть образ P, то и P есть (в данном преобразовании) образ P'. Это дает право называть точки P и P' *взаимно обратными* относительно окружности C. Инверсия превращает внутреннюю область окружности во внешнюю, и обратно: в самом деле, из неравенства $OP < r$ следует неравенство $OP' > r$ и, напротив, из неравенства $OP > r$ — неравенство $OP' < r$. Неподвижными точками плоскости являются точки самой окружности C.

Правило (1) не определяет никакого образа для центра O. Но ясно, что когда движущаяся точка P приближается к O, ее образ P' уходит неограниченно далеко. По этой причине иногда говорят, что при инверсии образом центра является *бесконечно удаленная точка.* Полезность этой терминологии вытекает из того обстоятельства, что она дает нам право утверждать, что инверсия устанавливает взаимно однозначное соответствие между всеми точками плоскости без исключения и их образами: каждая точка плоскости имеет один и только один образ и сама является образом одной и только одной точки. Отметим, что это последнее свойство принадлежит также и раньше приведенным примерам геометрических преобразований.

2. Свойства инверсии. Самое важное свойство инверсии заключается в том, что она *преобразует прямые линии и окружности в прямые линии и окружности.* Точнее, мы сейчас обнаружим, что в результате инверсии

а) прямая, проходящая через O, становится прямой, проходящей через O,

б) прямая, не проходящая через O, становится окружностью, проходящей через O,

в) окружность, проходящая через O, становится прямой, не проходящей через O,

г) окружность, не проходящая через O, становится окружностью, не проходящей через O.

Утверждение а) не требует доказательства, так как из самого определения инверсии ясно, что каждая точка на рассматриваемой прямой имеет в качестве образа другую точку на той же прямой, так что хотя отдельные точки на прямой перемещаются, но прямая в целом остается неизменной.

Докажем утверждение б). Из O опустим перпендикуляр на данную прямую L (рис. 39). Пусть A — основание этого перпендикуляра, A' — точка, обратная точке A. Возьмем произвольную точку P на L и обозначим через P' точку, ей обратную. Так как $OA \cdot OA' = OP \cdot OP' = r^2$, то отсюда следует, что

$$\frac{OA'}{OP'} = \frac{OP}{OA}.$$

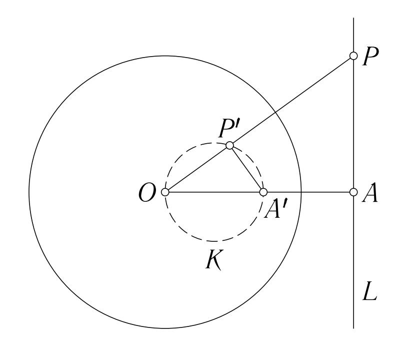


Рис. 39. Инверсия прямой относительно окружности

Поэтому треугольники $OP'A'$ и OAP подобны и, значит, угол $OP'A'$ прямой. В таком случае из теорем элементарной геометрии вытекает, что P' лежит на окружности K с диаметром OA'; эта окружность и является, следовательно, образом прямой L. Итак, утверждение б) доказано. Утверждение в) следует из того, что если образ L есть K, то образ K есть L.

Остается доказать утверждение г). Пусть K — окружность, не проходящая через O, с центром M и радиусом k (рис. 40). Чтобы получить ее образ, проведем через O прямую, пересекающую K в точках A и B, и затем посмотрим, как изменяются образы A' и B', когда направление прямой изменяется и она пересекает K самыми разнообразными способами.

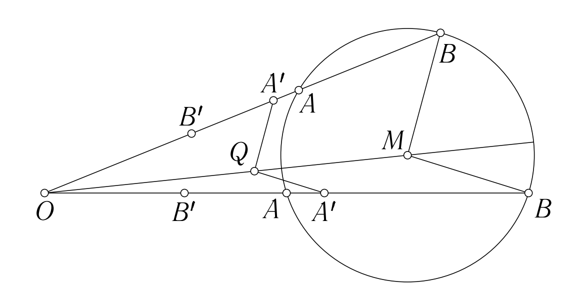


Рис. 40. Инверсия окружности

Обозначим расстояния OA, OB, OA', OB', OM через a, b, a', b', m, и пусть t есть длина касательной к K, проведенной из точки O. По определению инверсии, мы имеем $aa' = bb' = r^2$, а по элементарному геометрическому свойству окружности $ab = t^2$. Если разделим первые равенства на второе, то получим

$$\frac{a'}{b} = \frac{b'}{a} = \frac{r^2}{t^2} = c^2,$$

где c^2 зависит только от r и t и, значит, не зависит от положения точек A и B. Теперь проведем через A' прямую, параллельную BM; пусть Q есть точка ее пересечения с OM. Положим $OQ = q$, $A'Q = \rho$. Тогда

$$\frac{q}{m} = \frac{a'}{b} = \frac{\rho}{k},$$

или же

$$q = \frac{ma'}{b} = mc^2, \quad \rho = \frac{ka'}{b} = kc^2.$$

Это означает, что при всевозможных положениях A и B точка Q на прямой OM всегда будет одна и та же и что расстояние $A'Q$ также не будет меняться. Точно так же $B'Q = \rho$, так как $\frac{a'}{b} = \frac{b'}{a}$. Итак, образами точек A и B на K будут точки, расстояния которых от Q равны постоянной величине ρ, т. е. образ K есть окружность. Утверждение г) доказано.

3. Геометрическое построение обратных точек. Следующая теорема будет полезна в пункте 4 этого параграфа: *точка P', обратная данной точке P относительно окружности C, может быть построена геометрически с помощью одного только циркуля.* Рассмотрим сначала тот случай, когда точка P находится вне окружности C. Радиусом OP опишем круговую дугу с центром P, пересекающую C в точках R и S. Затем из этих точек как центров опишем круговые дуги радиусом r, равным радиусу круга C; эти дуги пересекутся в O и еще в точке P' на прямой OP. В равнобедренных треугольниках ORP и ORP'

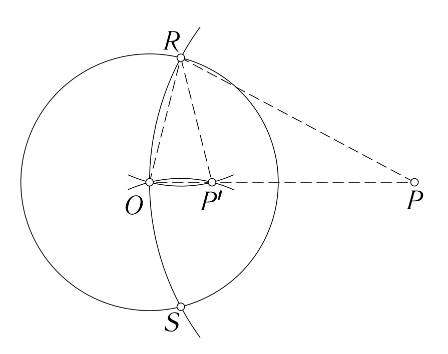


Рис. 41. Инверсия точки, внешней относительно окружности

$$\angle ORP = \angle POR = \angle OP'R,$$

так что треугольники подобны, и потому

$$\frac{OP}{OR} = \frac{OR}{OP'}, \quad \text{т. е.} \quad OP \cdot OP' = r^2.$$

Значит, P' есть искомая точка P.

Если данная точка P лежит внутри C, то построение и доказательство остаются в силе, лишь бы окружность радиуса OP с центром P пересекала окружность C в двух точках. Если же пересечений не получается, то можно редуцировать построение к предыдущему случаю посредством следующего простого приема.

Прежде всего заметим, что на прямой, соединяющей две данные точки A и O, можно с помощью одного циркуля построить такую точку C, что $AO = OC$. Для этого достаточно провести окружность с центром O и радиусом $r = AO$. Затем, начиная от точки A, отметить последовательно на этой окружности такие точки P, Q, C, что $AP = PQ = QC = r$. Тогда C есть как раз искомая точка: это ясно из того, что треугольники AOP, OPQ, OQC — равносторонние, так что угол между OA и OC содержит $180°$ и $OC = OQ = AO$. Повторяя указанную процедуру, мы имеем возможность отложить отрезок AO по прямой сколько угодно раз. Кстати, так как длина отрезка AO равна $r\sqrt{3}$ (как читатель проверит без всякого труда), то нам удалось построить $\sqrt{3}$, исходя из единичного отрезка, не пользуясь линейкой.

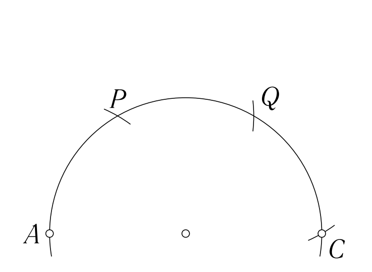


Рис. 42. Удвоение отрезка

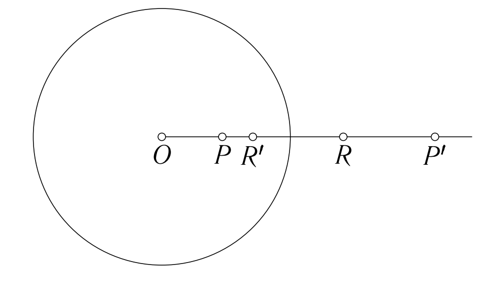


Рис. 43. Инверсия точки, внутренней относительно окружности

Теперь мы можем построить точку, обратную точке P относительно окружности C, как бы точка P ни была расположена внутри C. Прежде всего на прямой OP найдем такую точку R, что OR есть кратное OP, и вместе с тем R лежит уже вне C:

$$OR = n \cdot OP.$$

Для этого достаточно последовательно откладывать расстояние OP посредством циркуля, пока мы не выберемся из круга C. Затем с помощью уже известного построения найдем точку R', обратную точке R. Тогда будем иметь

$$r^2 = OR' \cdot OR = OR' \cdot (n \cdot OP) = (n \cdot OR') \cdot OP.$$

Останется построить точку P' по условию $OP' = n \cdot OR'$, и задача будет закончена.

4. Как разделить отрезок пополам и как найти центр данной окружности с помощью одного циркуля. После того как мы научились находить точку, обратную данной, можно с помощью одного циркуля выполнить дальнейшие интересные построения. Например, сейчас мы найдем середину отрезка, концы которого A и B заданы, с помощью

одного циркуля — не проводя самого отрезка. Вот решение этой задачи. Опишем окружность радиусом AB с центром B и на нем, отправляясь от A, как раньше, отмерим последовательно три дуги радиусом AB. Последняя точка C будет лежать на прямой AB, причем мы будем иметь: $AB = BC$. Затем опишем окружность радиуса AB с центром A и построим точку C', обратную точке C относительно этой окружности. Тогда получим:

$$AC' \cdot AC = AB^2,$$
$$AC' \cdot 2AB = AB^2,$$
$$2AC' = AB.$$

Значит, C' есть искомая середина отрезка.

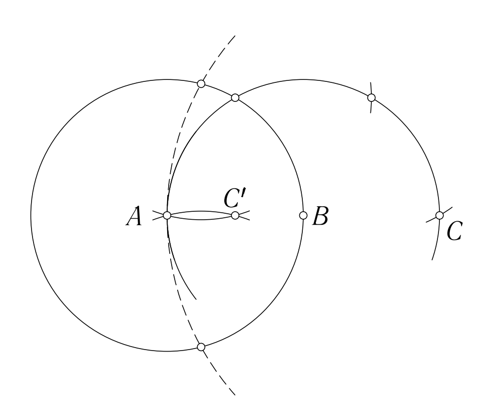


Рис. 44. Нахождение середины отрезка

Другое построение с помощью одного циркуля, также использующее обратные точки, заключается в нахождении центра данной окружности, когда начерчена только сама окружность, а центр неизвестен. Берем произвольную точку P на окружности и около нее как центра описываем круг произвольного радиуса, пересекающийся с данным кругом в точках R и S. Из этих последних точек как центров описываем дуги радиусом $RP = SP$, пересекающиеся, кроме точки P, еще в точке Q. Сравнивая то, что получилось, с рис. 41, мы видим, что неизвестный центр Q' есть точка, обратная точке Q относительно окружности с центром P, и Q' может быть, как мы видели, построена с помощью одного циркуля.

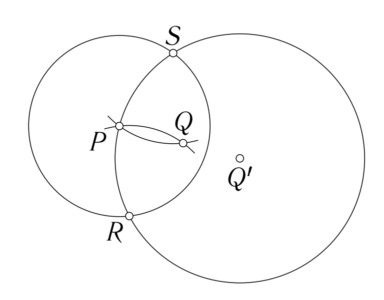


Рис. 45. Нахождение центра круга

§ 5. Построения с помощью других инструментов. Построения Маскерони с помощью одного циркуля

***1. Классическая конструкция, служащая для удвоения куба.** Мы рассматривали до сих пор только проблемы геометрических построений без использования иных инструментов, кроме циркуля и линейки. Если допускаются и другие инструменты, то, разумеется, разнообра-

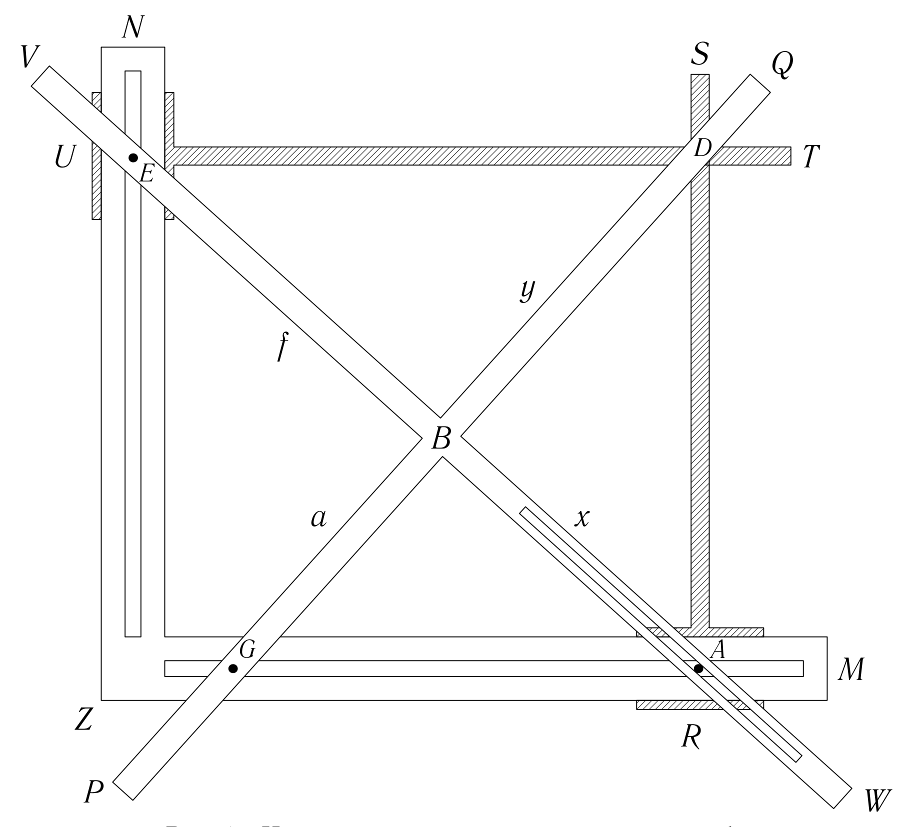


Рис. 46. Инструмент, служащий для удвоения куба

зие возможных построений сильно увеличивается. Следующий пример может служить образцом того, как греки решали проблему удвоения куба. Рассмотрим (рис. 46) жесткий прямой угол MZN и подвижной прямоугольный крест VW, PQ. Двум дополнительным стержням RS и TU предоставлена возможность скользить, оставаясь перпендикулярными к сторонам прямого угла. На кресте пусть выбраны фиксированные точки E и G, причем расстояния $GB = a$ и $BE = f$ заданы. Располагая крест таким образом, чтобы точки E и G соответственно лежали на NZ и MZ, и перемещая стержни TU и RS, можно весь аппарат привести в такое положение, чтобы лучевые перекладины креста BW, BQ, BV проходили через вершины A, D, E прямоугольника $ADEZ$. Указанное на чертеже расположение всегда возможно при условии $f > a$. Мы видим сразу, что $a : x = x : y = y : f$, откуда, в частности, если положено $f = 2a$, получается $x^3 = 2a^3$. Значит, x есть ребро куба, объем которого вдвое больше, чем объем куба с ребром a. Таким образом, поставленная задача

решена.

2. Построения с помощью одного циркуля. Если вполне естественно, что с допущением большего разнообразия инструментов оказывается возможным решать более обширное множество задач на построение, то можно было бы предвидеть, что, напротив, при ограничениях, налагаемых на инструменты, класс разрешимых задач будет суживаться. Тем более замечательным нужно считать открытие, сделанное итальянцем Маскерони (1750–1800): *все геометрические построения, выполнимые с помощью циркуля и линейки, могут быть выполнены с помощью одного только циркуля.* Следует, конечно, оговорить, что провести на самом деле прямую линию через две данные точки без линейки невозможно, так что это основное построение не покрывается теорией Маскерони. Вместо того приходится считать, что прямая задана, если заданы две ее точки. Но с помощью одного лишь циркуля удается найти точку пересечения двух прямых, заданных таким образом, или точку пересечения прямой с окружностью.

Вероятно, простейшим примером построения Маскерони является удвоение данного отрезка AB. Решение было уже дано на стр. 166. Далее, на стр. 167 мы научились делить данный отрезок пополам. Посмотрим теперь, как разделить пополам дугу окружности AB с центром O. Вот описание этого построения (рис. 47). Радиусом AO проводим две дуги с центрами A и B. От точки O откладываем на этих дугах две такие дуги OP и OQ, что $OP = OQ = AB$. Затем находим точку R пересечения дуги с центром P и радиусом PB и дуги с центром Q и радиусом QA. Наконец, взяв в качестве радиуса отрезок OR, опишем дугу с центром P или Q до пересечения с дугой AB — точка пересечения и является искомой средней точкой дуги AB. Доказательство предоставляем читателю в качестве упражнения.

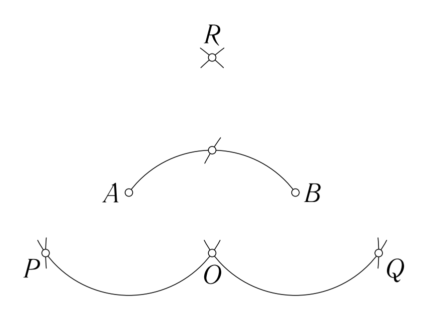


Рис. 47. Нахождение середины дуги без линейки

Было бы невозможно доказать основное утверждение Маскерони, указывая для каждого построения, выполнимого с помощью циркуля и линейки, как его можно выполнить с помощью одного циркуля: ведь возможных построений бесчисленное множество. Но мы достигнем той же цели, если установим, что каждое из следующих основных построений выполнимо с помощью одного циркуля:

1. Провести окружность, если заданы центр и радиус.

2. Найти точки пересечения двух окружностей.
3. Найти точки пересечения прямой и окружности.
4. Найти точку пересечения двух прямых.

Любое геометрическое построение (в обычном смысле, с допущением циркуля и линейки) составляется из выполнения конечной последовательности этих элементарных построений. Что первые два из них выполнимы с помощью одного циркуля, ясно непосредственно. Более трудные построения 3 и 4 выполняются с использованием свойств инверсии, рассмотренных в предыдущем пункте.

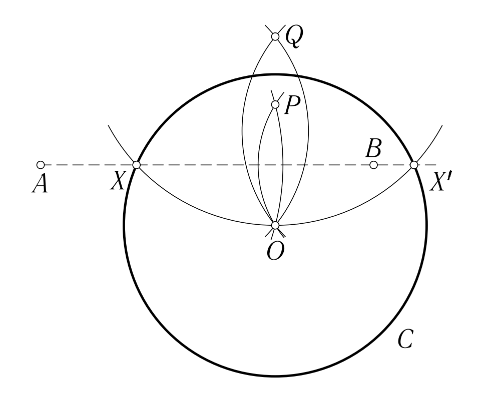


Рис. 48. Пересечение окружности и прямой, не проходящей через центр

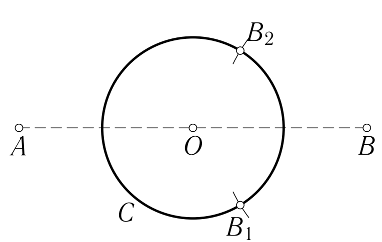


Рис. 49. Пересечение окружности и прямой, проходящей через центр

Обратимся к построению 3: найдем точки пересечения данной окружности C с прямой, проходящей через данные точки A и B. Проведем дуги с центрами A и B и радиусами, соответственно равными AO и BO; кроме точки O, они пересекутся в точке P. Затем построим точку Q, обратную точке P относительно окружности C (см. построение, описанное на стр. 167). Наконец, проведем окружность с центром Q и радиусом QO (она непременно пересечется с C): ее точки пересечения X и X' с окружностью C и будут искомыми. Для доказательства достаточно установить, что каждая из точек X и X' находится на одинаковых расстояниях от O и P (что касается точек A и B, то аналогичное их свойство сразу вытекает из построения). Действительно, достаточно сослаться на то обстоятельство, что точка, обратная точке Q, отстоит от точек X и X' на расстояние, равное радиусу окружности C (см. стр. 165). Стоит отметить, что окружность, проходящая через точки X, X' и O, является обратной прямой AB в инверсии относительно круга C, так как эта окружность и прямая AB пересекаются с C в одних и тех же точках. (При инверсии точки основной окружности остаются неподвижными.)

Указанное построение невыполнимо только в том случае, если прямая AB проходит через центр C. Но тогда точки пересечения могут быть найдены посредством построения, описанного на стр. 169, как середины дуг C, получающихся, когда мы проводим произвольную окружность с центром B, пересекающуюся с C в точках B_1 и B_2.

Метод проведения окружности, обратной прямой, соединяющей две данные точки, немедленно дает и построение, решающее задачу 4. Пусть прямые даны точками A, B и A', B' (рис. 50). Проведем произвольную окружность C и с помощью указанного выше метода построим окружности,
обратные прямым AB и $A'B'$. Эти окружности пересекаются в точке O и еще в одной точке Y. Точка X, обратная точке Y, и есть искомая точка пересечения: как ее построить — уже было разъяснено выше. Что X есть искомая точка, это ясно из того факта, что Y есть единственная точка, обратная точке, одновременно принадлежащей обеим прямым AB и $A'B'$; следовательно, точка X, обратная Y, должна лежать одновременно и на AB, и на $A'B'$.

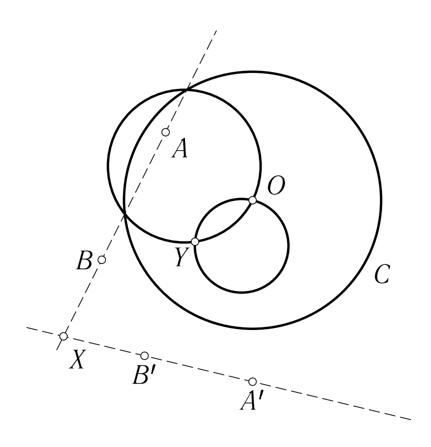


Рис. 50. Пересечение двух прямых

Этими двумя построениями заканчивается доказательство эквивалентности между построениями Маскерони, при которых разрешается пользоваться только циркулем, и обыкновенными геометрическими построениями с циркулем и линейкой.

Мы не заботились об изяществе решения отдельных проблем, нами здесь рассмотренных, так как нашей целью было выяснить внутренний смысл построений Маскерони. Но в качестве примера мы еще укажем пятиугольника; точнее говоря, речь идет о нахождении каких-то пяти точек на окружности, которые могут служить вершинами правильного вписанного пятиугольника.

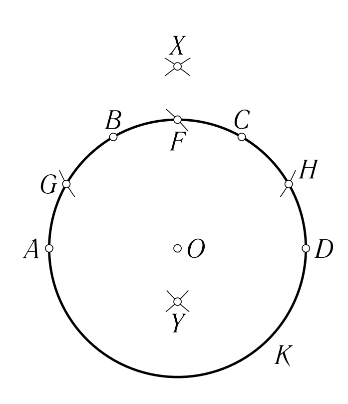


Рис. 51. Построение правильного пятиугольника

Пусть A — произвольная точка окружности K. Так как сторона правильного вписанного шестиугольника равна радиусу круга, то не представит труда отложить на K такие точки B, C, D, что $\smile AB = \smile BC = \smile CD = 60^\circ$ (рис. 51). Проведем дуги с центрами A и D радиусом, рав-

ным AC; пусть они пересекаются в точке X. Тогда, если O есть центр K, дуга с центром A и радиусом OX пересечет K в точке F, являющейся серединой дуги BC (см. стр. 169). Затем радиусом, равным радиусу K, опишем дуги с центром F, пересекающиеся с K в точках G и H. Пусть Y есть точка, расстояния которой от точек G и H равны OX и которая отделена от X центром O. В таком случае отрезок AY как раз и есть сторона искомого пятиугольника. Доказательство предоставляется читателю в качестве упражнения. Интересно отметить, что при построении используются только три различных радиуса.

В 1928 г. датский математик Ельмслев нашел в книжной лавке в Копенгагене экземпляр книги под названием Euclides Danicus, опубликованной в 1672 г. никому не известным автором Г. Мором. По титульному листу можно было сделать заключение, что это — просто один из вариантов евклидовых «Начал», снабженный, может быть, редакторским комментарием. Но при внимательном рассмотрении оказалось, что в ней содержится полное решение проблемы Маскерони, найденное задолго до Маскерони.

Упражнения. В дальнейшем дается описание построений Мора. Проверьте их правильность. Почему можно утверждать, что они решают проблему Маскерони?

1) К отрезку AB длины p восставите перпендикуляр BC. (*Указание*: продолжите AB до точки D таким образом, что $AB = BD$. Проведите произвольным радиусом дуги с центрами A и D и таким образом определите C.)

2) В плоскости даны как угодно расположенные отрезки длины p и q, причем $p > q$. Постройте с помощью 1) отрезок длины $x = \sqrt{p^2 - q^2}$.

3) По заданному отрезку a постройте отрезок $a\sqrt{2}$. (*Указание*: обратите внимание, что $(a\sqrt{2})^2 = (a\sqrt{3})^2 - a^2$.)

4) По данным отрезкам p и q постройте отрезок $x = \sqrt{p^2 + q^2}$. (*Указание*: примите во внимание, что $x^2 = 2p^2 - (p^2 - q^2)$.) Придумайте сами аналогичные построения.

5) Пользуясь предыдущими результатами, постройте отрезки $p + q$ и $p - q$, предполагая, что отрезки длины p и q заданы как-то на плоскости.

6) Проверьте и постарайтесь обосновать следующее построение середины M данного отрезка AB длины a. На продолжении отрезка AB найдем такие точки C и D, что $CA = AB = BD$. Построим равносторонний треугольник ECD согласно условию $EC = ED = 2a$ и определим M как пересечение окружностей с диаметрами EC и ED.

7) Найдите прямоугольную проекцию точки A на отрезок BC.

8) Найдите x по условию $x : a = p : q$, где a, p и q — данные отрезки.

9) Найдите $x = ab$, где a и b — данные отрезки.

Вдохновляясь результатами Маскерони, Якоб Штейнер (1796–1863) предпринял попытку исследования построений, выполнимых с помощью одной только линейки. Конечно, одна только линейка не выводит за

пределы данного числового поля, и потому она недостаточна для выполнения всех геометрических построений в классическом их понимании. Но тем более замечательны результаты, полученные Штейнером при введенном им ограничении — пользоваться циркулем только один раз. Он доказал, что все построения на плоскости, выполнимые с помощью циркуля и линейки, выполнимы также с помощью одной линейки при условии, что задан единственный неподвижный круг вместе с центром. Эти построения подразумевают применение проективных методов и будут описаны позднее (см. стр. 217).

* Без круга, и притом с центром, обойтись нельзя. Например, если дан круг, но не указан его центр, то найти центр с помощью одной линейки невозможно. Мы сейчас докажем это, ссылаясь, однако, на факт, который будет установлен позднее (см. стр. 240): существует такое преобразование плоскости самой в себя, что а) заданная окружность остается неподвижной, б) всякая прямая линия переходит в прямую, в) центр неподвижной окружности не остается неподвижным, а смещается. Само существование такого преобразования свидетельствует о невозможности построить центр данной окружности, пользуясь одной линейкой. В самом деле, какова бы ни была процедура построения, она сводится к ряду отдельных этапов, заключающихся в проведении прямых линий и нахождении их пересечений друг с другом или с данной окружностью. Представим себе теперь, что вся фигура в целом — окружность и все прямые, проведенные по линейке при выполнении построения центра — подвергнута преобразованию, существование которого мы здесь допустили. Тогда ясно, что фигура, полученная после преобразования, также удовлетворяла бы всем требованиям построения; но указываемое этой фигурой построение приводило бы к точке, отличной от центра данной окружности. Значит, построение, о котором идет речь, невозможно.

3. Черчение с помощью различных механических приспособлений. Механические кривые. Циклоиды. Изобретение различных механизмов, предназначенных для того, чтобы чертить различные кривые, помимо окружности и прямой линии, чрезвычайно расширяет область фигур, допускающих построение. Например, если имеется инструмент, позволяющий чертить гиперболы $xy = k$, и другой инструмент, вычерчивающий параболы $y = ax^2 + bx + c$, то любая проблема, приводящая к кубическому уравнению

$$ax^3 + bx^2 + cx = k, \tag{1}$$

может быть решена конструктивно, с помощью только этих инструментов. В самом деле, решение уравнения (1) равносильно решению системы

$$xy = k, \quad y = ax^2 + bx + c; \tag{2}$$

точнее, корни уравнения (1) являются x-координатами точек пересечения гиперболы и параболы, представляемых уравнениями (2). Таким

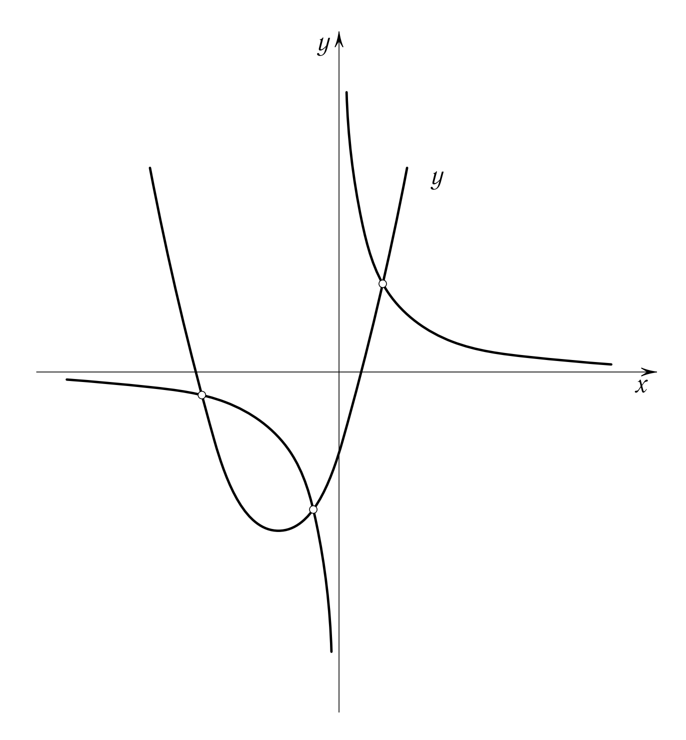


Рис. 52. Графическое решение кубического уравнения

образом, решения уравнения (1) допускают построение, если разрешается пользоваться инструментами, с помощью которых можно начертить кривые (2).

Уже математикам древности были известны многие интересные кривые, которые могут быть определены и начерчены с помощью простых механических приспособлений. Среди таких «механических» кривых особенно видное место занимают *циклоиды.* Птолемей (около 200 года до нашей эры), обнаруживая необычайную проницательность, сумел использовать эти кривые для описания планетных движений.

Циклоида самого простого вида представляет собой траекторию движения точки P, фиксированной на окружности диска, катящегося без скольжения по прямой линии. На рис. 53 изображены четыре положения точки P в различные моменты времени. По форме циклоида напоминает ряд арок, опирающихся на горизонтальную прямую.

Разновидности этой кривой получаются, если возьмем точку P или внутри диска (как на спице колеса), или на продолжении радиуса за пределы диска.

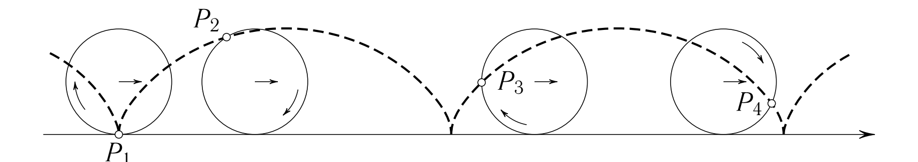


Рис. 53. Циклоида

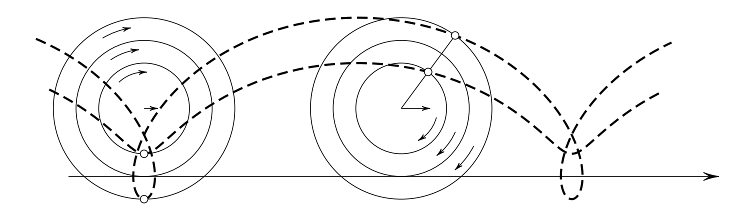


Рис. 54. Циклоиды общего вида

Эти две кривые показаны на рис. 54.

Дальнейшие разновидности циклоиды возникают, когда наш диск катится не по прямой, а по дуге окружности. Если при этом катящийся диск с радиусом r остается все время касающимся *изнутри* той большой окружности C радиуса R, по которой он катится, то траектория точки, фиксированной на окружности диска, называется *гипоциклоидой*.

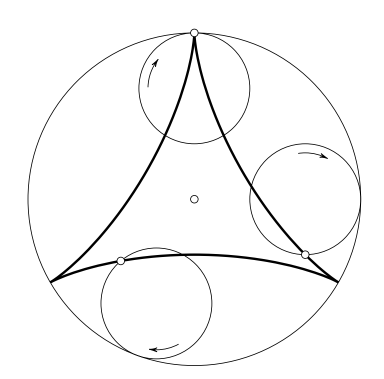


Рис. 55. Трехрогая гипоциклоида

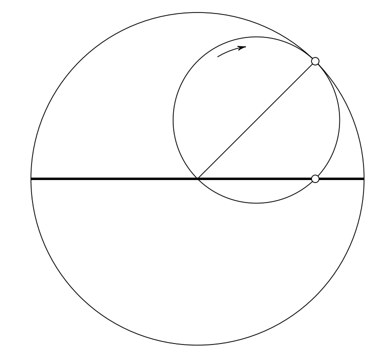


Рис. 56. Прямолинейное движение при качении круга по кругу двойного радиуса

Когда диск прокатывается по всей окружности C ровно один раз, то точка P возвращается в исходное положение только в том случае, если радиус C является кратным радиуса c. На рис. 55 изображена замкнутая гипоциклоида, соответствующая предположению $R = 3r$. В более общем

случае, если $R = \frac{m}{n} r$, то гипоциклоида замкнется после того, как диск c прокатится по окружности C ровно n раз, и будет состоять из m арок. Заслуживает особого упоминания случай $R = 2r$. Любая точка P на окружности диска будет описывать в этом случае один из диаметров большой окружности C (рис. 56). Предоставляем читателю доказать это в качестве задачи.

Еще один тип циклоид получается, когда диск c катится по окружности C, касаясь ее все время *извне*. Получающиеся при этом кривые носят название *эпициклоид*.

***4. Шарнирные механизмы. Инверсоры Поселье и Гарта.** Оставим на время в стороне вопрос о циклоидах (они появятся еще раз в этой книге — довольно неожиданно) и обратимся к иным методам механического воспроизведения кривых линий. Мы займемся сейчас *шарнирными механизмами*.

Механизм этого типа представляет собой систему сочлененных между собой твердых стержней, обладающих такой степенью свободы, чтобы каждая его точка была способна описывать определенную кривую. Циркуль также является простейшим шарнирным механизмом, по существу состоящим из одного стержня с закрепленным концом.

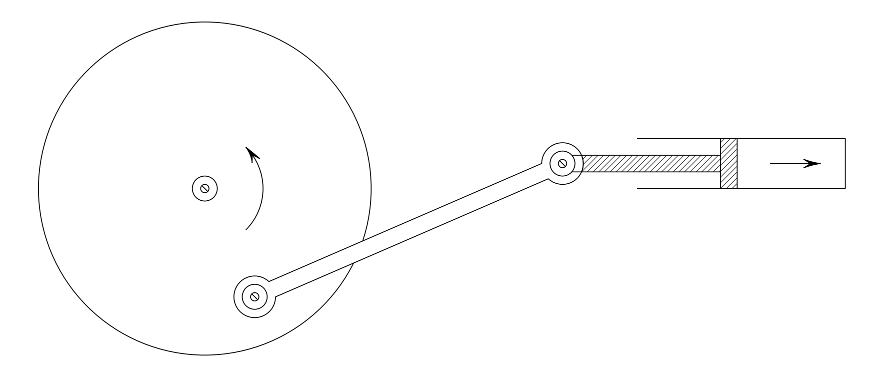


Рис. 57. Преобразование прямолинейного движения во вращательное

Шарнирные механизмы издавна находят себе применение как составные части машин. Одним из самых знаменитых (в историческом отношении) примеров является так называемый «параллелограмм Уатта». Это приспособление было изобретено Джемсом Уаттом при решении следующей проблемы: как связать поршень с точкой махового колеса таким образом, чтобы вращение колеса сообщало поршню прямолинейное движение? Решение, данное Уаттом, было лишь приближенным, и, несмотря на усилия многих первоклассных математиков, проблема конструирования механизма, сообщающего точке в *точности* прямолиней-

ное движение, долгое время оставалась нерешенной. Было даже сделано предположение, что такой механизм неосуществим: это было как раз тогда, когда всякого рода «доказательства невозможности» привлекли к себе всеобщее внимание. Тем большее изумление было вызвано в кругах математиков, когда французский морской офицер Поселье (в 1864 г.) все же изобрел несложный механизм, действительно разрешающий проблему в положительном смысле. В связи с введением в употребление хорошо действующих смазочных веществ техническая проблема потеряла свое значение для паровых машин.

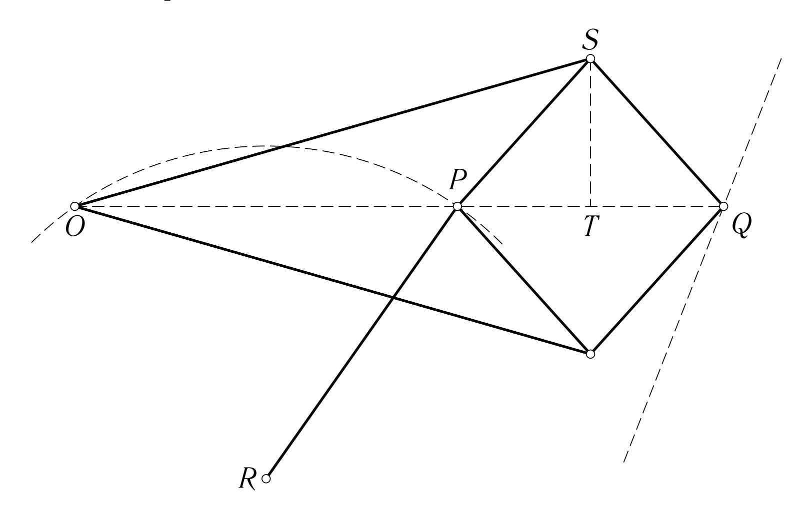


Рис. 58. Инверсор Поселье, преобразующий вращательное движение в прямолинейное

Назначение механизма Поселье заключается в том, чтобы превращать круговое движение в прямолинейное. В основе этого механизма лежит теория инверсии, изложенная в § 4. Как видно из рис. 58, механизм состоит из семи жестких стержней, два из них — длины t, четыре — длины s и один — произвольной длины. Точки O и R закреплены и расположены таким образом, что $OR = PR$. Весь аппарат может быть приведен в движение, будучи подчинен указанным условиям. Мы сейчас убедимся, что, *когда точка P описывает дугу окружности с центром R и радиусом RP, точка Q описывает прямолинейный отрезок.* Обозначая основание перпендикуляра, опущенного из точки S на прямую OPQ, через T, мы замечаем, что

$$OP \cdot OQ = (OT - PT) \cdot (OT + PT) = OT^2 - PT^2 = \\ = (OT^2 + ST^2) - (RT^2 + ST^2) = t^2 - s^2. \quad (3)$$

Величина $t^2 - s^2$ постоянная; положим $t^2 - s^2 = r^2$. Так как $OP \cdot OQ =$

r^2, то точки P и Q взаимно обратные относительно окружности с центром O и радиусом r. В то время как P описывает дугу окружности, проходящей через O, Q описывает кривую, обратную этой дуге. Но кривая, обратная окружности, проходящей через O, есть, как мы видели, не что иное, как прямая линия. Итак, траектория точки Q есть прямая, и инверсор Поселье чертит эту прямую без линейки.

Другой механизм, решающий ту же проблему, есть инверсор Гарта. Он состоит всего лишь из пяти стержней, сочленение которых показано на рис. 59. Здесь $AB = CD$, $BC = AD$. Через O, P и Q обозначены точки, соответственно зафиксированные на стержнях AB, AD и CB, притом таким образом, что $\frac{AO}{OB} = \frac{AP}{PD} = \frac{CQ}{QB} = \frac{m}{n}$. Точки O и S закреплены на плоскости неподвижно, с соблюдением условия $OS = PS$. Больше связей нет, и механизм способен двигаться. Очевидно, прямая AC всегда

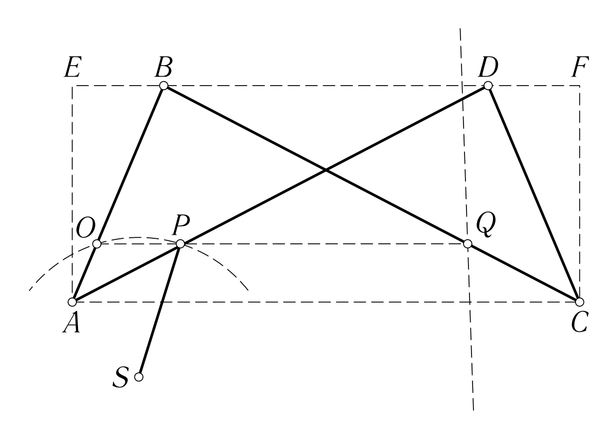


Рис. 59. Инверсор Гарта

параллельна прямой BD. В таком случае точки O, P и Q лежат на одной прямой, и прямая OP параллельна прямой AC. Проведем перпендикуляры AE и CF к прямой BD. Мы имеем

$$AC \cdot BD = EF \cdot BD = (ED + EB) \cdot (ED - EB) = ED^2 - EB^2.$$

Но $ED^2 + AE^2 = AD^2$ и $EB^2 + AE^2 = AB^2$. Значит, $ED^2 - EB^2 = AD^2 - AB^2$. Далее,

$$\frac{OP}{BD} = \frac{AO}{AB} = \frac{m}{m+n} \quad \text{и} \quad \frac{OQ}{AC} = \frac{OB}{AB} = \frac{n}{m+n}.$$

Следовательно,

$$OP \cdot OQ = \frac{mn}{(m+n)^2} \cdot BD \cdot AC = \frac{mn}{(m+n)^2} \cdot (AD^2 - AB^2).$$

Последняя полученная величина не изменяется при движении механизма. Поэтому точки P и Q являются взаимно обратными относительно

некоторого круга с центром O. При движении механизма точка P описывает окружность с центром S, проходящую через O; значит, обратная точка Q описывает прямую линию.

Можно построить — по крайней мере теоретически — другие шарнирные механизмы, которые будут чертить эллипсы, гиперболы и даже любую наперед заданную алгебраическую кривую $f(x, y) = 0$, какова бы ни была ее степень.

§ 6. Еще об инверсии и ее применениях

1. Инвариантность углов. Семейства окружностей. Хотя круговая инверсия есть преобразование, довольно резко меняющее внешний вид геометрических фигур, все же весьма замечательным является то обстоятельство, что вновь получаемые фигуры сохраняют некоторые свойства первоначальных фигур. Эти свойства, не теряющиеся при преобразовании, называются *инвариантными*. Мы уже знаем, что при инверсии окружность или прямая переходит в окружность или прямую. Прибавим теперь еще одно важное свойство инверсии: *угол между двумя прямыми или кривыми при инверсии не изменяется.* Говоря подробнее, это означает, что инверсия преобразовывает две пересекающиеся кривые в две другие кривые, которые пересекаются под тем же углом. Под углом между кривыми подразумевается угол между их касательными.

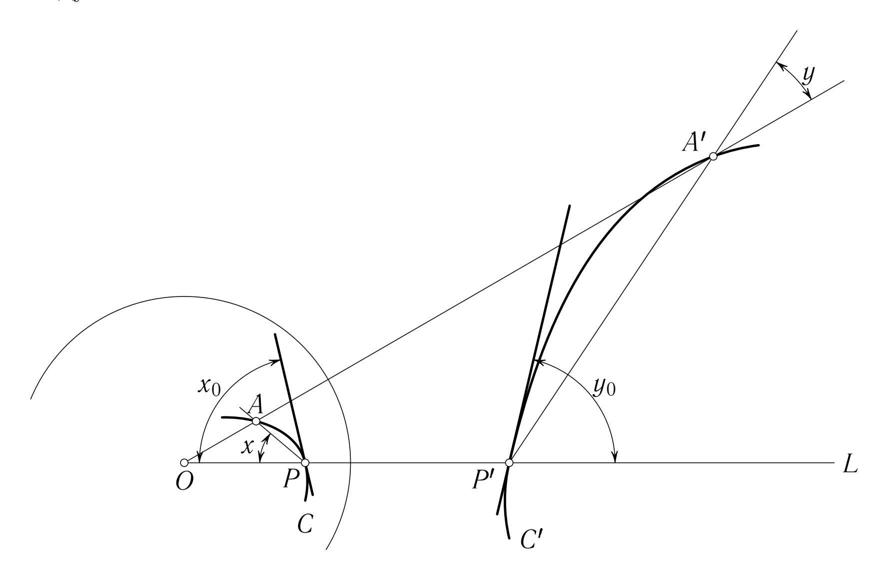


Рис. 60. Инвариантность углов при инверсии

Доказательство получается при рассмотрении рис. 60, где имеется

в виду частный случай пересечения в точке P произвольной кривой C с прямолинейным отрезком OL, проведенным из центра инверсии O. Кривая C', обратная кривой C, пересекается с OL в точке P', обратной P, так как P', так же как и P, лежит на OL. Покажем, что угол x_0 между OL и касательной к C в точке P по величине равен углу y_0 между OL и касательной к C' в точке P'. Для этого возьмем точку A на кривой C вблизи P и проведем секущую AP.

Точка, обратная A, есть A'; так как она находится на прямой OA и на кривой C', то является их точкой пересечения. Проведем также секущую $A'P'$. По определению инверсии, $r^2 = OP \cdot OP' = OA \cdot OA'$, или же

$$\frac{OP}{OA} = \frac{OA'}{OP'},$$

т. е. треугольники OAP и $OA'P'$ подобны. Значит, угол x равен углу $OA'P'$, который мы обозначим через y. Последний шаг в нашем рассуждении заключается в том, чтобы заставить точку A приближаться по кривой C к точке P. При этом секущая AP переходит в касательную к кривой C в точке P, и угол x стремится к x_0. В то же время A' будет приближаться к P' и прямая $A'P'$ перейдет в касательную к кривой C' в точке P', а угол y будет стремиться к y_0. Так как при всяком положении точки A мы имеем равенство $x = y$, то оно сохранится и в пределе $x_0 = y_0$.

Наше доказательство еще не закончено, так как мы рассмотрели пока только случай пересечения кривой C с прямой, проходящей через центр O. Но рассмотреть общий случай пересечения двух произвольных кривых C и C^* теперь уже совсем легко. Пусть эти кривые пересекаются в точке P и образуют между собой угол z. Тогда прямая OPP' делит этот угол на два угла, из которых каждый в отдельности не изменяется при инверсии.

Следовало бы оговорить, что, хотя инверсия не изменяет *величины* угла, она, однако, изменяет *направление* его отсчета: если вообразим, что при постоянном увеличении угла x_0 одна сторона его неподвижна, а другая вращается против часовой стрелки, то подвижная сторона соответствующего «обратного» угла вращается по часовой стрелке.

Частным следствием инвариантности углов при инверсии является то, что две ортогональные (т. е. пересекающиеся под прямым углом) окружности или прямые после инверсии сохраняют это свойство, и если две окружности взаимно касаются («пересекаются под углом, равным нулю»), то касаются и обратные им окружности.

Рассмотрим семейство окружностей, проходящих через центр инверсии O и еще через одну и ту же неподвижную точку плоскости A. Мы знаем (§ 4, пункт 2), что это семейство преобразуется в семейство прямых, проходящих через точку A', являющуюся образом A. В то же время семейство окружностей, ортогональных первоначальному семейству,

превращается в семейство окружностей, ортогональных упомянутому семейству прямых. (На рис. 61 ортогональные семейства изображены пунктиром.) Внешне семейство прямых, проходящих через одну и ту же точку, мало напоминает семейство окружностей, но эти семейства связаны теснейшим образом — с точки зрения теории инверсии они, так сказать, вполне эквивалентны.

Вот другой пример того, к каким результатам приводит инверсия. Пусть дано семейство окружностей, проходящих через центр инверсии и имеющих в этой точке общую касательную. После инверсии получается семейство параллельных прямых. Действительно, так как окружности проходят через O, то они превращаются в прямые, и так как окружности не имеют точек пересечения кроме O, то получаемые прямые параллельны.

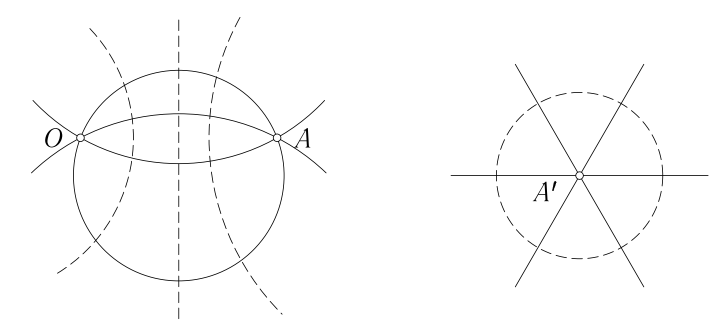


Рис. 61. Преобразование двух систем ортогональных окружностей с помощью инверсии

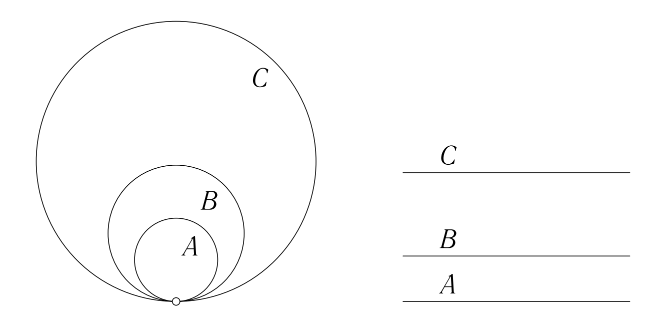


Рис. 62. Преобразование касающихся окружностей в параллельные прямые

2. Применение к проблеме Аполлония. Прекрасной иллюстрацией того, насколько полезна теория инверсии, является следующее простое геометрическое решение проблемы Аполлония. При инверсии отно-

сительно какого бы то ни было центра проблема Аполлония для трех данных окружностей трансформируется в соответствующую проблему для трех других окружностей: пусть читатель внимательно продумает, почему это так.

Отсюда легко понять, что если проблема решена для некоторой тройки окружностей, то тем самым ее можно считать решенной и для всякой тройки окружностей, которая из первой тройки может быть получена путем инверсии. Мы сумеем использовать это обстоятельство, выбирая из всевозможных «эквивалентных» троек такую, для которой проблема решается особенно просто.

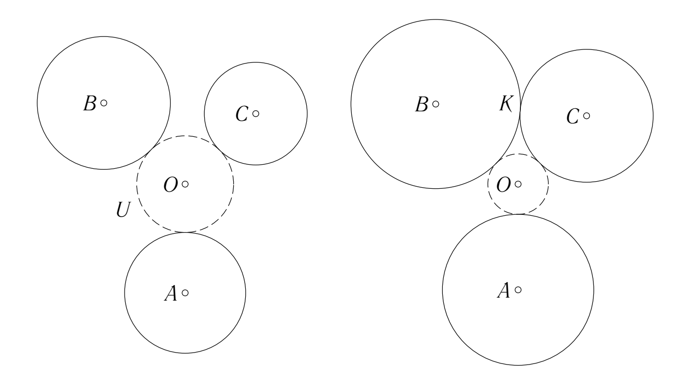


Рис. 63. Подготовка построения, решающего проблему Аполлония

Предположим для определенности, что три данные окружности с центрами A, B, C взаимно не пересекаются и лежат каждая вне двух других, и допустим, что речь идет о нахождении окружности U с центром O и радиусом ρ, касающейся трех данных окружностей внешним образом. Заметим, что если мы увеличим радиус всех трех данных окружностей на одну и ту же величину d, то окружность с центром O и радиусом $\rho - d$, очевидно, был бы решением видоизмененной таким образом проблемы. Пользуясь этим обстоятельством, увеличим радиусы данных окружностей на такую величину, чтобы две из трех окружностей оказались взаимно касающимися в некоторой точке, которую обозначим K (рис. 63). Затем произведем инверсию всей фигуры относительно какой-нибудь окружности с центром K. Окружности с центрами B и C станут параллельными прямыми b и c, а третья окружность превратится в некоторую окружность a (рис. 64). Мы уже знаем, что a, b, c могут

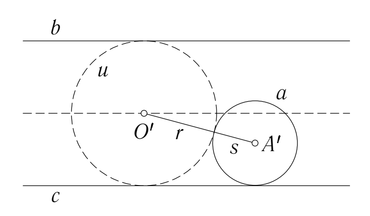


Рис. 64. Решение проблемы Аполлония

быть построены с помощью циркуля и линейки. Что касается искомой окружности U, то она преобразуется в окружность u, касающуюся прямых b, c и окружности a. Ее радиус r, очевидно, должен равняться половине расстояния между прямыми b и c; центр же ее O' должен совпадать с одной из точек пересечения средней линии между b и c с окружностью, концентрической окружности a, но имеющей радиус на r больший. Остается применить обратную инверсию к окружности u, и тогда получим искомую аполлониеву окружность U.

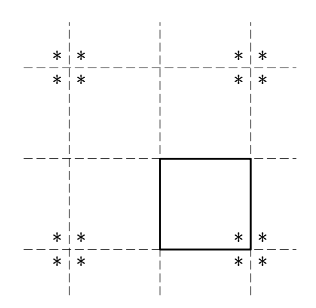


Рис. 65. Повторное отражение относительно прямолинейных стен

3. Повторные отражения. Каждому из нас приходилось наблюдать странные явления отражения, возникающие, если имеется более одного зеркала. Если четыре стены прямоугольной комнаты представляют собой идеальные зеркала, ни в малой степени не поглощающие света, то находящаяся в этой комнате освещенная точка создает бесконечное множество отражений, по одному на каждую из прямоугольных комнат, возникающих из первой посредством отражений (рис. 65). При менее правильной форме соединения зеркал, например при трех зеркалах, создается более сложная система отражений. Получающуюся

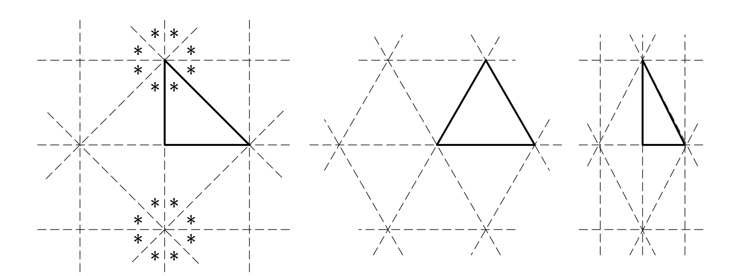


Рис. 66. Правильные системы треугольных зеркал

конфигурацию легко описать только в том случае, если отраженные треугольники, не перекрывая друг друга, полностью покрывают плоскость. Таким свойством обладают только прямоугольный равнобедренный треугольник, равносторонний треугольник и прямоугольный треугольник, представляющий собою половину равностороннего (рис. 66).

Еще более курьезные обстоятельства возникают, если мы станем рассматривать повторные инверсии относительно пары окружностей. Поместившись между двумя концентрическими сферическими зер-

калами, мы увидали бы бесчисленное множество концентрических отражений. Одна последовательность отражений уходила бы в бесконечность, другая — сосредоточивалась бы около центра. Случай двух окружностей, расположенных одна вне другой, несколько сложнее: окружности и их отражения последовательно отражаются одна в другой, уменьшаясь после каждого отражения и теснясь к двум предельным точкам, по одной в каждой из данных окружностей. (Эти точки обладают свойством взаимной обратности относительно каждой из данных окружностей.) Все это показано на рис. 67. Что получится в случае трех кругов, об этом читатель может составить впечатление, взглянув на узор, изображенный на рис. 68.

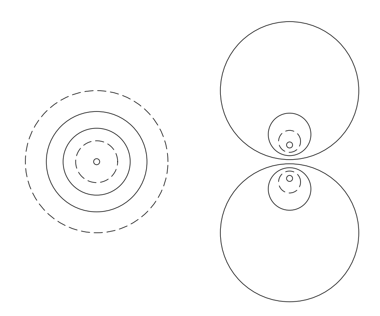


Рис. 67. Повторное отражение относительно двух сферических дуг

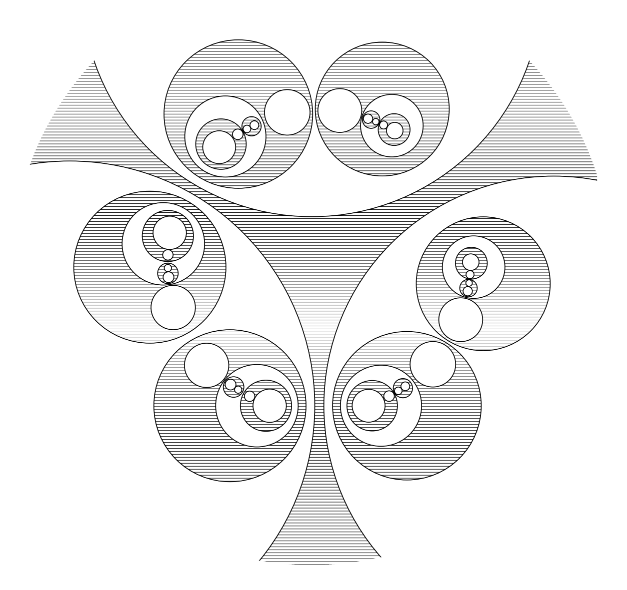


Рис. 68. Отражение относительно трех сферических зеркал

ГЛАВА IV

Проективная геометрия. Аксиоматика. Неевклидовы геометрии

§ 1. Введение

1. Классификация геометрических свойств. Инвариантность при преобразованиях. В геометрии рассматриваются свойства фигур на плоскости и в пространстве. Эти свойства столь многочисленны и столь разнообразны, что необходим какой-то принцип классификации для того, чтобы внести порядок в обширное собрание накопленных знаний. Можно было бы, например, положить в основу классификации метод, применяемый при выводе получаемых утверждений. С этой точки зрения обыкновенно различаются «синтетические» и «аналитические» процедуры. Синтетические доказательства существенно связаны с классическим аксиоматическим методом, идущим от Евклида: рассуждение строится на чисто геометрической основе, независимо от средств алгебры и концепции числового континуума, и все теоремы выводятся формально логическим путем, исходя из некоторого числа начальных положений, называемых аксиомами или постулатами. Другой метод подразумевает введение числовой координатной системы и использует технический аппарат алгебры. Этот метод произвел глубокие изменения в самой математической науке, слив в одно органическое целое геометрию, анализ и алгебру.

В этой главе, однако, нас будет интересовать не столько классификация методов, сколько классификации содержания, т. е. сами по себе утверждения теорем, а не способы их доказательства. В элементарной геометрии плоскости резко различаются две группы теорем; в одних идет речь о равенстве фигур, об измерении отрезков и углов, в других — о подобии фигур, для которого существенно измерение углов, но не отрезков. Указанное различие не столь уж существенно, так как длины отрезков и величины углов довольно тесно связаны между собой и разделять их — несколько искусственно. (Изучение этой связи составляет главным образом предмет тригонометрии.) Отметим иную сторону дела. В элементарной геометрии мы имеем дело с *величинами*: отрезками, углами, площадями. Две фигуры там считаются эквивалент-

ными, если они *конгруэнтны*, т. е. могут быть переведены одна в другую посредством движения — преобразования, меняющего только положение фигуры, но не числовые значения величин, с ней связанных. Возникает вопрос: является ли значение величин — и вместе с тем конгруэнтность или подобие фигур — чем-то существенно неизменным в геометрии? Или же имеются иные, более глубоко лежащие свойства геометрических фигур, которые сохраняются также и при преобразованиях более общего типа, чем движения? Мы увидим, что такие свойства существуют.

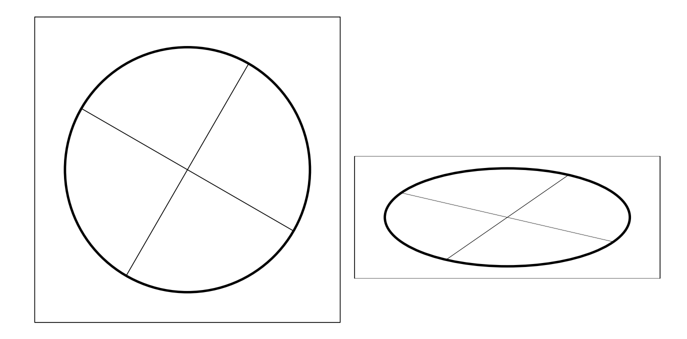


Рис. 69. Сжатие окружности

Представим себе, что на прямоугольной доске, изготовленной из мягкого эластичного материала, нарисован круг с парой взаимно перпендикулярных диаметров (рис. 69). Если мы положим эту доску в тиски и сожмем до половины ее первоначальной ширины, то окружность превратится в эллипс и углы между диаметрами уже не будут прямыми. Окружность обладает тем свойством, что все ее точки находятся на одном и том же расстоянии от центра, но эллипс таким свойством не обладает. Могло бы показаться, что сжатие уничтожает все геометрические свойства первоначальной конфигурации. Но и это предположение далеко от истины: например, утверждение, что центр делит диаметры пополам, одинаково справедливо и для окружности, и для эллипса; в данном случае мы встречаемся с таким свойством фигуры, которое сохраняется при весьма резком изменении в размерах ее отдельных элементов. Сделанные замечания наводят на мысль о возможности классифицировать теоремы, относящиеся к той или иной геометрической фигуре, в зависимости от того, сохраняют ли они силу или теряют ее при равномерном сжатии (или растяжении). Можно поставить и более общий вопрос, исходя из некоторого данного класса преобразований фигуры (такого рода классы, например, порождаются совокупностью всех движений, или сжатий, или, скажем, круговых инверсий и т. д.); можно поинтересоваться тем, какие свойства фигуры остаются неизменными, когда фигура подвергается различным преобразованиям данного класса.

Система теорем, утверждающих такие свойства, составляет *геометрию рассматриваемого класса преобразований.* Идея классификации различных отраслей геометрии в соответствии с классами преобразований принадлежит Феликсу Клейну (1849–1925); она была высказана им в 1872 г. в его знаменитом выступлении, получившем широкую известность под названием «Эрлангенской программы». С тех пор эта идея оказала решающее влияние на направление многих геометрических исследований.

В главе V нам представится случай установить весьма удивительное обстоятельство, заключающееся в том, что некоторые свойства геометрических фигур заложены настолько глубоко, что не исчезают даже после совершенно произвольных деформаций: так, фигуры, нарисованные на куске резины, не потеряют кое-каких характеристических черт при самых разнообразных и самых резких деформациях. Но в настоящей главе мы займемся теми свойствами, которые сохраняются, «инвариантны» при некотором специальном классе преобразований, более широком, чем весьма ограниченный класс движений, но более узком, чем самый общий класс произвольных деформаций. Мы говорим о классе «проективных преобразований».

2. Проективные преобразования. Изучение относящихся сюда геометрических свойств было выдвинуто перед математиками в давнее время проблемами *перспективы*, которые изучались художниками, в том числе Леонардо да Винчи и Альбрехтом Дюрером. Изображение, создаваемое художником, следует рассматривать как проекцию оригинала на плоскость картины, причем центр проекции помещается в глазу художника. При проектировании — в зависимости от относительных положений различных изображаемых объектов — длины отрезков и углы неизбежно подвергаются искажениям. И тем не менее на картине обычно не представляет труда распознать геометрическую структуру оригинала. Как объяснить это обстоятельство? Нельзя объяснить иначе, как указав на наличие геометрических свойств, «инвариантных относительно проектирования»,— свойств, сохраняющихся на картине и делающих возможным узнавание нарисованного оригинала. Отыскание и анализ этих свойств составляют предмет проективной геометрии.

Совершенно ясно, что в этой отрасли геометрии не содержится положительных утверждений, относящихся к длинам отдельных отрезков или к величинам отдельных углов; не идет речь и о равенстве фигур. Некоторые изолированные факты, касающиеся проективных свойств, были известны уже в XVII в., а иногда, как в случае «теоремы Менелая», даже в древности. Но систематические исследования в области проективной геометрии развернулись впервые лишь к концу XVIII столетия, когда знаменитая École Polytechnique в Париже открыла новую страницу в истории математики, в частности геометрии. Эта школа, создан-

ная Французской революцией, подготовила большое число офицеров, оказавших на военной службе выдающиеся услуги своей республике. В числе ее питомцев был Жан-Виктор Понселе (1788–1867), написавший свой «Трактат о проективных свойствах фигур» в 1813 г., будучи в плену в России.

В XIX в. под влиянием Штейнера, Штаудта, Шаля и других проективная геометрия стала одним из излюбленных предметов математических исследований. Своей популярностью она обязана отчасти присущей ей особенной эстетической привлекательности, отчасти же способности проливать свет на геометрическую науку в целом, а также глубокой внутренней связи с неевклидовой геометрией и с алгеброй.

§ 2. Основные понятия

1. Группа проективных преобразований. Прежде всего определим класс, или «группу»[1], проективных преобразований. Пусть в пространстве заданы две плоскости π и π', параллельные или непараллельные между собой. Мы выполняем *центральную проекцию* π на π' с данным центром O, не лежащим ни на π, ни на π', сопоставляя каждой точке P плоскости π такую точку P' плоскости π', что P и P' лежат на одной и той же прямой, проходящей через O. Аналогично мы выполняем подобным же образом *параллельную проекцию*, предполагая, что проектирующие прямые параллельны между собой. Точно так же определяется проекция прямой или кривой линии l в плоскости π на некоторую линию l' в плоскости π', причем и в этом случае проекция может быть центральной или параллельной.

Всякое отображение одной фигуры на другую, получающееся посредством проектирования (центрального или параллельного) или же посредством конечной последовательности таких проектирований, называется *проективным преобразованием*[2]. *Проективная геометрия* плоскости или прямой составляется из системы геометрических теорем, сохраня-

---

1 Термин «группа» в применении к классу преобразований подразумевает, что последовательное выполнение двух преобразований из рассматриваемого класса есть также преобразование этого класса и что преобразование, «обратное» по отношению к преобразованию из рассматриваемого класса, также принадлежит этому классу. Групповые свойства математических операций играли и продолжают играть очень большую роль во многих областях, однако по отношению к геометрии значение понятия «группы» в свое время, возможно, было несколько преувеличено.

2 Если две фигуры связаны только одним проектированием, то говорят обычно, что они перспективны. Таким образом, если сказано, что фигура F в результате проективного преобразования переходит в фигуру F', то это значит, что или фигуры F и F' перспективны, или же можно указать последовательность таких фигур F, F_1, F_2, ..., F_n, F', что любые две рядом стоящие в ней фигуры перспективны.

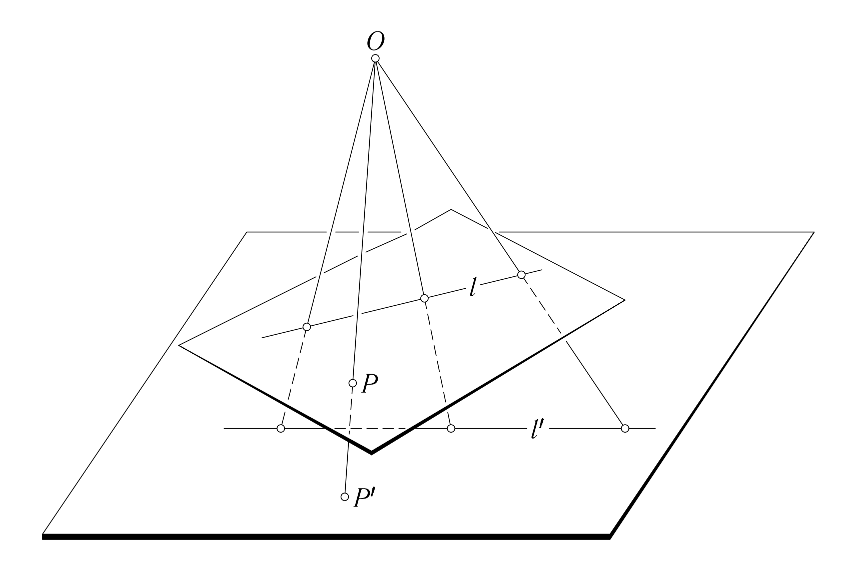


Рис. 70. Центральная проекция

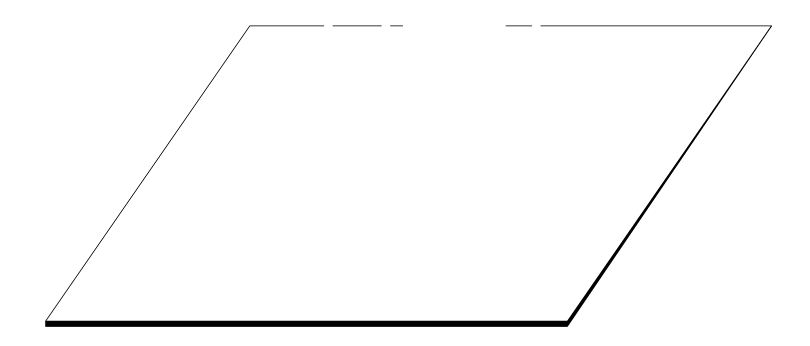


Рис. 71. Параллельная проекция

ющихся при произвольных проективных преобразованиях соответствующих фигур. Проективной геометрии противопоставляется *метрическая геометрия*, которая понимается как система теорем, устанавливающих связи между величинами в рассматриваемых фигурах, инвариантные только относительно класса движений.

Некоторые проективные свойства можно формулировать непосредственно. Точка, разумеется, проектируется в точку. Далее, *прямая линия проектируется в прямую*: в самом деле, если прямая l в плоскости π проектируется на плоскость π', то линия пересечения l' плоскости π с плоскостью, проходящей через O и l, — обязательно прямая[1]. Если точка A и прямая l инцидентны[2], то точка A' и прямая l', возникающие из них при проективном преобразовании, также инцидентны. Другими словами, *инцидентность точки и прямой есть свойство, инвариантное относительно группы проективных преобразований*. Из этого обстоятельства вытекает ряд простых, но весьма важных следствий. Если три точки (или более трех точек) *коллинеарны*, т. е. инцидентны с одной и той же прямой, то их отображения также коллинеарны. Аналогично, если в плоскости π три прямые (или более трех прямых) *конкуррентны*, т. е. инцидентны с одной и той же точкой, то их отображения — также конкуррентные прямые. В то время как эти простые свойства — инцидентность, коллинеарность, конкуррентность — являются *проективными свойствами* (т. е. свойствами, инвариантными относительно проективных преобразований), величины отрезков и углов, а также и отношения этих величин, вообще говоря, изменяются при проектировании. Равнобедренные или равносторонние треугольники могут, например, спроектироваться на треугольники с тремя различными сторонами. Отсюда следует, что хотя понятие «треугольник» принадлежит проективной геометрии, понятие «равносторонний треугольник» ей не принадлежит, а принадлежит только метрической геометрии.

2. Теорема Дезарга. Одним из самых ранних открытий в области проективной геометрии является замечательная теорема Дезарга (1593–1662): *если на плоскости два треугольника ABC и $A'B'C'$ расположены таким образом, что прямые, соединяющие соответственные вершины, конкуррентны, то три точки, в которых пересекаются, будучи продолжены, три соответственные стороны, коллинеарны.* Эта теорема здесь иллюстрируется чертежом (рис. 72), но пусть читатель проверит ее справедливость, экспериментируя на самостоятельно построенных

---

1 За исключением того случая, когда прямая OP (или плоскость, проходящая через O и l) оказывается параллельной плоскости π. Такие исключения будут устранены в § 4.

2 Точка и прямая называются инцидентными, если прямая проходит через точку или точка лежит на прямой. Этот «нейтральный» термин подчеркивает взаимность рассматриваемого отношения.

чертежах. Доказательство теоремы не является тривиальным, несмотря на всю простоту чертежа, состоящего только из прямых линий. Теорема явственно принадлежит проективной геометрии, так как при проектировании рассматриваемый чертеж не теряет свойств, упомянутых в теореме. В дальнейшем мы еще вернемся к этой теореме (стр. 207). В на-

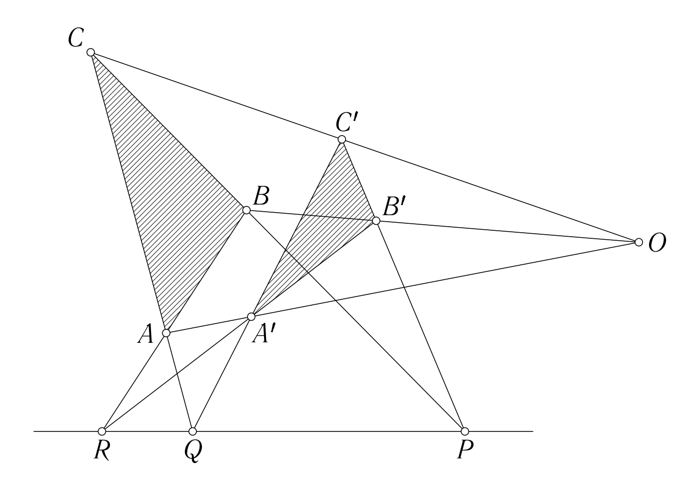


Рис. 72. Конфигурация Дезарга на плоскости

стоящий же момент мы хотели бы привлечь внимание читателя к тому любопытному обстоятельству, что теорема Дезарга справедлива также и в том предположении, что рассматриваемые треугольники расположены в двух *различных* (непараллельных) плоскостях и что подобного рода «трехмерная», или «пространственная» теорема Дезарга доказывается без малейших затруднений. По предположению, прямые AA', BB' и CC' пересекаются в одной и той же точке O (рис. 73). В таком случае прямые AB и $A'B'$ лежат в одной плоскости и, значит, пересекаются в некоторой точке R; пусть, таким же образом, AC и $A'C'$ пересекаются в точке Q, а BC и $B'C'$ — в точке P. Так как точки P, Q и R находятся на продолжениях сторон треугольников ABC и $A'B'C'$, то все они лежат в плоскости каждого из этих треугольников и потому — на прямой пересечения этих двух плоскостей. Значит P, Q и R коллинеарны, что и требовалось доказать.

Это простое доказательство наводит на мысль, что можно попытаться доказать «двумерную» теорему Дезарга, так сказать, с помощью перехода к пределу, постепенно сплющивая всю пространственную конструкцию таким образом, чтобы две плоскости в пределе совпали в одну, и в этой последней, вместе с другими, оказалась и точка O. Выполнить, однако, указанный предельный переход не так просто, так как прямая пересечения PQR при совмещении плоскостей не определяется

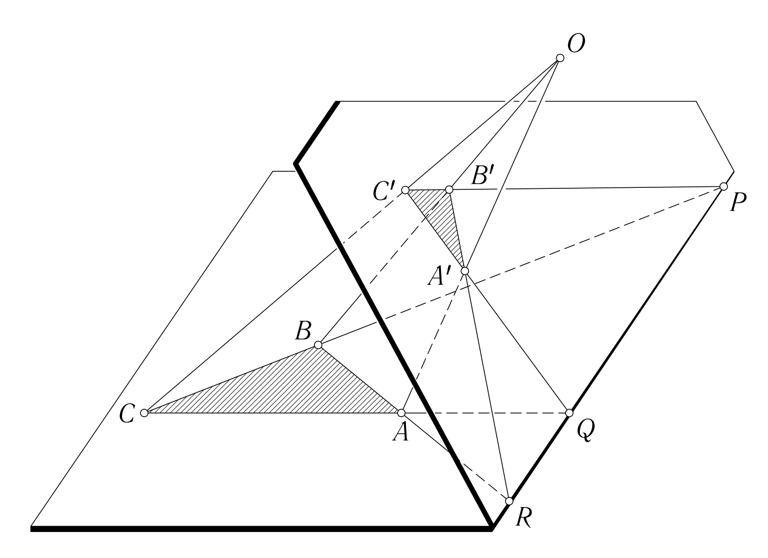


Рис. 73. Конфигурация Дезарга в пространстве

однозначно.

Тем не менее конфигурация, изображенная на рис. 72, может быть истолкована как перспективное изображение пространственной конфигурации рис. 73, и это обстоятельство можно использовать при доказательстве «плоской» теоремы.

Есть существенное различие между теоремой Дезарга на плоскости и в пространстве. Наше доказательство, относящееся к случаю трех измерений, опиралось только на геометрические соображения, относящиеся к инцидентности точек, прямых и плоскостей. Можно показать, что доказательство двумерной теоремы — *при дополнительном требовании не выходить из данной плоскости* — неизбежно должно опираться на некоторые свойства подобных фигур, принадлежащих уже не проективной, а метрической геометрии.

Теорема, обратная теореме Дезарга, утверждает, что если три точки, в которых пересекаются соответственные стороны, коллинеарны, то прямые, соединяющие соответственные вершины, конкуррентны. Доказательство этой теоремы — в случае, когда треугольники лежат в непараллельных плоскостях, — предоставляется читателю в качестве упражнения.

§ 3. Двойное отношение

1. Определение и доказательство инвариантности. Если длина отрезка прямой представляет собой своего рода ключ к метрической геометрии, то существует и в проективной геометрии одно основное понятие, с помощью которого могут быть выражены все отличительные проективные свойства фигур.

Предположим, что три точки A, B и C расположены на одной прямой. Проектирование, вообще говоря, изменяет не только расстояния AB и BC, но и их отношение $\frac{AB}{BC}$. В самом деле, *любые* три точки A, B, C на прямой l могут быть переведены в *любые* три точки A', B', C' на прямой l' посредством двух последовательно производимых проектирований. Чтобы в этом убедиться, станем вращать прямую l' около точки C', пока она не примет положения l'', параллельного l (рис. 74). Затем, проектируя l на l'' параллельно прямой CC', получим три точки A'', B'' и C'' ($\equiv C'$). Прямые $A'A''$ и $B'B''$ пересекутся в точке O, которую мы изберем в качестве центра второй проекции. Последовательно выполненные указанные две проекции дают требуемый результат[1].

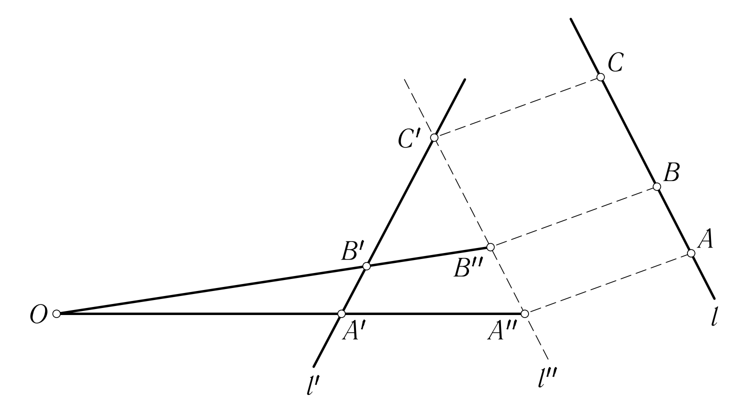


Рис. 74. Проектирование трех точек

Из доказанного вытекает, что никакая величина, определяемая только тремя точками на прямой, не может быть инвариантной при проектировании. Но — в этом заключается решающее открытие в проективной геометрии — если на прямой дано *четыре* точки A, B, C, D, которые при проектировании переходят в точки A', B', C', D' другой прямой, то некоторая величина, называемая *двойным отношением* этих четырех точек, при проектировании не изменяет числового значения. В этом заключено математическое свойство системы четырех точек на прямой, которое носит инвариантный характер и которое можно обнаружить во всякой проекции рассматриваемой прямой. Двойное отношение не есть ни расстояние, ни отношение расстояний, а *отношение двух таких отношений*: если мы составим отношения

$$\frac{CA}{CB} \quad \text{и} \quad \frac{DA}{DB},$$

[1] Подумайте, что делать, если прямые $A'A''$ и $B'B''$ параллельны.

то их отношение

$$x = \frac{CA}{CB} : \frac{DA}{DB}$$

по определению есть двойное отношение четырех точек A, B, C, D, взятых в указанном выше порядке.

Убедимся теперь, что *двойное отношение четырех точек инвариантно при проектировании*, т. е. что если A, B, C, D и A', B', C', D' — две четверки точек на двух прямых и между ними установлено проективное соответствие, то тогда справедливо равенство

$$\frac{CA}{CB} : \frac{DA}{DB} = \frac{C'A'}{C'B'} : \frac{D'A'}{D'B'}.$$

Доказательство вполне элементарно. Вспомним, что площадь треугольника равна половине произведения основания на высоту и, с другой стороны, равна половине произведения двух сторон на синус заключенного между ними угла. Тогда получим (рис. 75):

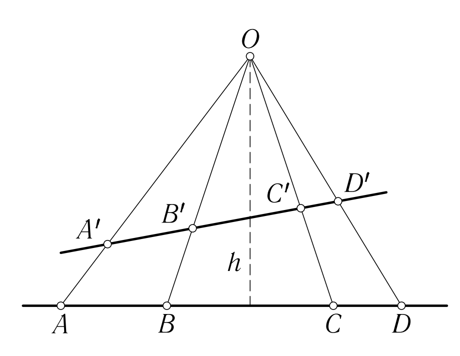


Рис. 75. Инвариантность двойного отношения при центральном проектировании

$$\text{площадь } OCA = \frac{1}{2}\,h \cdot CA = \frac{1}{2}\,OA \cdot OC \sin\angle COA,$$

$$\text{площадь } OCB = \frac{1}{2}\,h \cdot CB = \frac{1}{2}\,OB \cdot OC \sin\angle COB,$$

$$\text{площадь } ODA = \frac{1}{2}\,h \cdot DA = \frac{1}{2}\,OA \cdot OD \sin\angle DOA,$$

$$\text{площадь } ODB = \frac{1}{2}\,h \cdot DB = \frac{1}{2}\,OB \cdot OD \sin\angle DOB.$$

Отсюда следует:

$$\frac{CA}{CB} : \frac{DA}{DB} = \frac{CA}{CB} \cdot \frac{DB}{DA} = \frac{OA \cdot OC \sin\angle COA}{OB \cdot OC \sin\angle COB} \cdot \frac{OB \cdot OD \sin\angle DOB}{OA \cdot OD \sin\angle DOA} = \frac{\sin\angle COA}{\sin\angle COB} \cdot \frac{\sin\angle DOB}{\sin\angle DOA}.$$

Таким образом, двойное отношение точек A, B, C, D зависит только от углов, образованных в точке O отрезками OA, OB, OC, OD. Так как эти углы — одни и те же, каковы бы ни были четыре точки A', B', C', D', в которые при проектировании переходят A, B, C, D, то ясно, что двойное отношение не изменяется при проектировании.

Что двойное отношение не изменяется при параллельном проектировании, следует из элементарных свойств подобных треугольников. Доказательство предоставляется читателю в качестве упражнения (рис. 76).

До сих пор, говоря о двойном отношении четырех точек A, B, C, D, расположенных на прямой l, мы предполагали, что это отношение составлено из положительных отрезков. Целесообразно видоизменить это определение следующим образом. Примем одно из двух направлений прямой l за положительное и условимся, что все отрезки, отсчитываемые в этом направлении, будут считаться положительными, а отрезки, отсчитываемые в противоположном направлении, — отрицательными. Теперь определим двойное отношение точек A, B, C, D (взятых в указанном порядке) согласно формуле

$$(ABCD) = \frac{CA}{CB} : \frac{DA}{DB},$$

причем знаки чисел CA, CB, DA, DB берутся в соответствии с указанным выше условием. Так как при изменении направления на прямой l, принятого за положительное, меняются только знаки всех четырех отрезков, то значение $(ABCD)$ не зависит от выбора направления. Легко понять, что $(ABCD)$ имеет отрицательный или положительный знак, смотря по тому, разделена ли пара точек A, B парой точек C, D или не разделена. Так как свойство «разделяться» инвариантно относительно проектирования, то понимаемое в новом смысле (как величина, способная иметь тот или иной знак) двойное отношение $(ABCD)$ также инвариантно. Выберем начальную точку O на прямой l и сопоставим каждой точке на прямой l в качестве координаты x ее расстояние от O, взятое с надлежащим знаком; тогда, обозначая координаты A, B, C, D соответственно через x_1, x_2, x_3, x_4, получим формулу

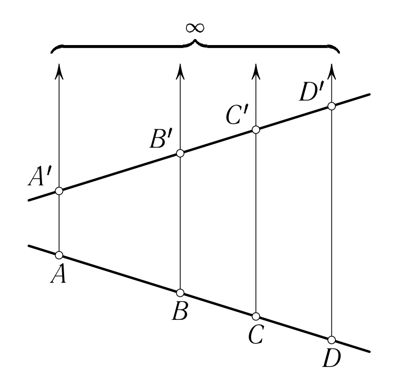


Рис. 76. Инвариантность двойного отношения при параллельном проектировании

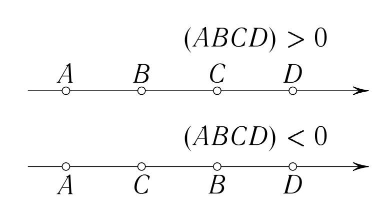


Рис. 77. Знак двойного отношения

$$(ABCD) = \frac{CA}{CB} : \frac{DA}{DB} = \frac{x_3 - x_1}{x_3 - x_2} : \frac{x_4 - x_1}{x_4 - x_2} = \frac{x_3 - x_1}{x_3 - x_2} \cdot \frac{x_4 - x_2}{x_4 - x_1}.$$

Если $(ABCD) = -1$, так что $\frac{CA}{CB} = -\frac{DA}{DB}$, то точки C и D делят отрезок AB внутренне и внешне в одном и том же отношении. В этом случае принято говорить, что C и D делят отрезок AB *гармонически*, и каждая из точек C и D считается *гармонически сопряженной* с другой точкой относительно пары точек A, B. Если $(ABCD) = 1$, то точки C и D (или A и B) совпадают.

Необходимо не упустить из виду, что при определении двойного отношения $(ABCD)$ существенную роль играет порядок, в котором берутся точки A, B, C, D. Например, если $(ABCD) = \lambda$, то двойное отношение $(BACD)$ равно $\frac{1}{\lambda}$, тогда как $(DACB) = 1 - \lambda$, в чем читатель убедится без труда. Четыре точки A, B, C, D могут быть переставлены между собой $4 \cdot 3 \cdot 2 \cdot 1 = 24$ различными способами, и каждой перестановке соответствует некоторое значение двойного отношения. Некоторым перестановкам соответствует то же числовое значение двойного отношения, что и начальной перестановке A, B, C, D; например, $(ABCD) = (BADC)$. Читателю предоставляется в качестве упражнения доказать, что при 24 возможных перестановках четырех точек получается всего лишь шесть различных значений двойного отношения, а именно

$$\lambda, \quad 1-\lambda, \quad \frac{1}{\lambda}, \quad \frac{\lambda-1}{\lambda}, \quad \frac{1}{1-\lambda}, \quad \frac{\lambda}{\lambda-1}.$$

Эти шесть величин, вообще говоря, различны, но при некоторых значениях λ могут и совпадать по две, например при значении $\lambda = -1$ в случае гармонического деления.

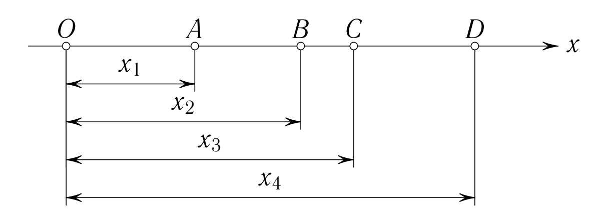


Рис. 78. Координатное выражение для двойного отношения

Мы можем также определить *двойные отношения четырех компланарных* (т. е. лежащих в одной плоскости) *и конкуррентных прямых* 1, 2, 3, 4, как двойное отношение четырех точек пересечения этих прямых с некоторой прямой, лежащей в той же плоскости. Положение этой пятой прямой несущественно вследствие инвариантности двойного отношения при проектировании. Эквивалентным этому определению является следующее:

$$(1234) = \pm\frac{\sin(1,\,3)}{\sin(2,\,3)} : \frac{\sin(1,\,4)}{\sin(2,\,4)},$$

где нужно взять знак плюс, если пара прямых 1, 2 не разделяется парой 3, 4, и знак минус, если разделяется. (В этой формуле (1, 3), например, обозначает угол между прямыми 1 и 3.) Наконец, можно определить двойное отношение четырех *коаксиальных плоскостей* (четырех плоскостей, пересекающихся по одной прямой, или «оси»). Если некоторая прямая пересекает плоскости в четырех точках, то двойное отношение этих точек всегда будет иметь одно и то же значение, независимо от

выбора прямой (доказательство предлагается в качестве упражнения). Таким образом, полученное значение можно назвать двойным отношением рассматриваемых четырех плоскостей. Иначе, можно назвать двойным отношением четырех коаксиальных плоскостей двойное отношение четырех прямых, по которым они пересекаются произвольной пятой плоскостью (рис. 79).

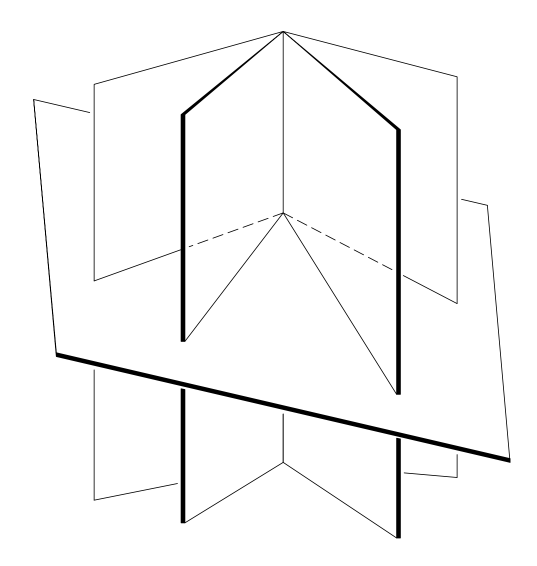


Рис. 79. Двойное отношение четырех плоскостей

Понятие двойного отношения четырех плоскостей побуждает поставить вопрос о том, нельзя ли дать определение проективного преобразования *трехмерного* пространства самого на себя. Определение с помощью центральной проекции, очевидно, не обобщается непосредственно от случая двух измерений на случай трех измерений. Но можно доказать, что каждое непрерывное преобразование плоскости самой на себя, взаимно однозначно переводящее точки в точки и прямые в прямые, есть проективное. Это обстоятельство наводит на мысль ввести следующее определение для случая трех измерений: проективным преобразованием пространства называется непрерывное взаимно однозначное преобразование, переводящее прямые линии в прямые линии. Можно показать, что такие преобразования оставляют значения двойных отношений неизменными.

Добавим к предыдущему еще кое-какие замечания. Пусть на прямой даны три различные точки A, B, C с координатами x_1, x_2, x_3. Требуется найти четвертую точку D таким образом, чтобы удовлетворялось равенство $(ABCD) = \lambda$, где λ задано. (Частный случай, когда $\lambda = -1$

и задача заключается в построении четвертой гармонической точки, будет подробно рассмотрен в следующем пункте.) Вообще говоря, задача имеет одно и только одно решение; действительно, если x — координата искомой точки D, то уравнение

$$\frac{x_3 - x_1}{x_3 - x_2} \cdot \frac{x - x_2}{x - x_1} = \lambda$$

имеет ровно одно решение. Считая x_1, x_2 и x_3 заданными и полагая ради краткости $\frac{x_3 - x_1}{x_3 - x_2} = k$, мы придадим решению вид

$$x = \frac{kx_2 - \lambda x_1}{k - \lambda}.$$

Например, если точки A, B, C находятся на равных расстояниях друг от друга и имеют соответственно координаты $x_1 = 0$, $x_2 = d$, $x_3 = 2d$, то тогда $k = \frac{2d - 0}{2d - d} = 2$ и $x = \frac{2d}{2 - \lambda}$.

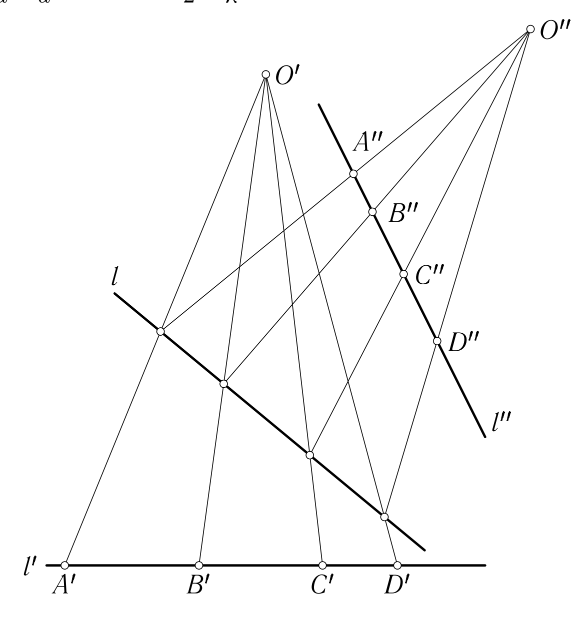


Рис. 80. Проективное соответствие между точками двух прямых

Если прямая l спроектирована из двух различных центров O' и O'' на две различные прямые l' и l'', то получается соответствие $P \longleftrightarrow P'$ между точками прямых l и l' и соответствие $P \longleftrightarrow P''$ между точками прямых l и l''. Этим устанавливается соответствие $P' \longleftrightarrow P''$ между точками прямых l' и l'', и притом такое, что каждые четыре точки A', B', C', D' на l' имеют то же самое двойное отношение, что и соответствующие точки A'', B'', C'', D'' на l''. Всякое взаимно однозначное соответствие между точками двух прямых, обладающее этим свойством, называется

проективным соответствием, независимо от того, каким способом это соответствие установлено.

Упражнения. 1) Докажите, что если даны две прямые вместе с проективным соответствием, установленным между ними, то можно подвергнуть одну из прямых такому параллельному перенесению, что заданное соответствие будет получаться посредством простой проекции. (*Указание*: совместите какую-нибудь пару взаимно соответствующих точек на данных прямых.)

2) Пользуясь предыдущим результатом, покажите, что если между точками двух прямых l и l' установлено соответствие посредством конечного числа последовательных проектирований на различные промежуточные прямые при произвольных центрах проекций, то тот же результат может быть получен посредством всего лишь *двух* проектирований.

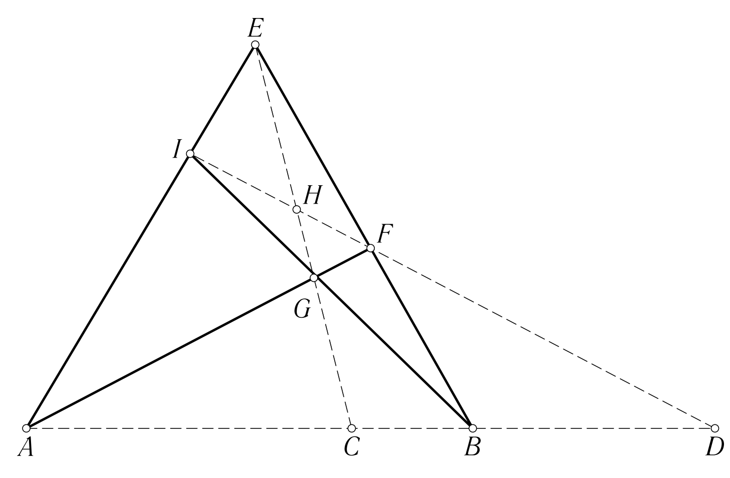


Рис. 81. Полный четырехсторонник

2. Применение к полному четырехстороннику. В качестве интересного применения инвариантности двойного отношения мы докажем одну простую, но важную теорему проективной геометрии. Речь идет о *полном четырехстороннике* — фигуре, образованной произвольными четырьмя прямыми, из которых никакие три не являются конкуррентными, и шестью точками их пересечения. На рис. 81 названные четыре прямые суть AE, BE, BI, AF. Прямые AB, EG и IF являются *диагоналями* четырехсторонника. Рассмотрим одну из диагоналей, например AB, и отметим на ней точки C и D, где она пересекается с двумя другими диагоналями. Тогда теорема утверждает существование равенства $(ABCD) = -1$; словами это выражается так: *точки пересечения одной диагонали с двумя другими делят отрезок между вершинами четырехсторонника гармонически.* Для доказательства достаточно об-

ратить внимание на то, что

$$x = (ABCD) = (IFHD) \qquad \text{(проектируем из } E\text{)},$$
$$(IFHD) = (BACD) \qquad \text{(проектируем из } G\text{)}.$$

Как нам известно,

$$(BACD) = \frac{1}{(ABCD)};$$

таким образом, $x = \frac{1}{x}$, $x^2 = 1$, $x = \pm 1$. Но так как C, D разделяют A, B, то двойное отношение x отрицательно и потому оно должно быть равно именно -1, что мы и хотели доказать.

Полученное замечательное свойство полного четырехсторонника дает нам возможность с помощью одной лишь линейки построить точку, гармонически сопряженную с точкой C относительно пары A, B (если A, B, C коллинеарны). Нужно только, выбрав произвольную точку E вне данной прямой, провести прямые EA, EB, EC; затем, взяв произвольно точку G на EC, провести прямые AD и BD, пересекающие EB и EA, скажем, в точках F и I; провести, наконец, прямую IF, которая и пересечет исходную прямую в искомой точке D.

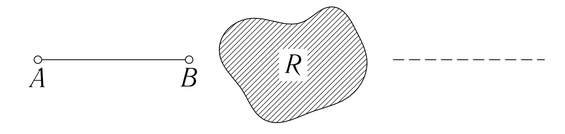


Рис. 82. Проведение прямой через препятствие

Задача. На плоскости задан отрезок AB и область R, как показано на рис. 82. Желательно продолжить прямую AB вправо от R. Как это можно сделать с помощью одной линейки и при условии, чтобы в процессе построения не покрывать линейкой никакой части области R? (*Указание*: выберите на отрезке AB две произвольные точки C и C', затем постройте сопряженные с ними гармонические D и D' относительно пары точек A, B; при построении воспользуйтесь четыре раза теоремой о полном четырехсторонника.)

§ 4. Параллельность и бесконечность

1. «Идеальные» бесконечно удаленные точки. Внимательное рассмотрение изложенного в предыдущем параграфе обнаруживает, что во многих случаях приведенная аргументация теряет силу — именно, тогда, когда прямые, точка пересечения которых нужна для построения, оказываются параллельными. Например, построение четвертой гармонической точки D становится невыполнимым, если прямая IF параллельна AB. Геометрические рассуждения на каждом шагу затруднены тем обстоятельством, что параллельные прямые не имеют общей точки,

и потому всякий раз, когда речь идет о пересечении прямых, приходится отдельно рассматривать и особо оговаривать случай параллелизма. С другой стороны, если производится проектирование, мы вынуждены различать и трактовать независимо рядом с центральной также и параллельную проекцию. Если бы из такого положения не было выхода, то проективная геометрия, будучи вынуждена вникать в детальное исследование каждого встречающегося исключения и особого случая, неизбежно была бы чрезвычайно усложнена. Все это побуждает искать выхода в ином направлении, именно, на пути *такого обобщения основных понятий, которое устраняло бы возможные исключения.*

Тут нам поможет геометрическая интерпретация; мы видим, что если прямая, пересекающая другую прямую, медленно вращается, приближаясь к положению параллельности, то точка пересечения двух прямых неограниченно удаляется. Это дает повод к наивному утверждению: две прямые пересекаются «в бесконечно удаленной точке». Подобного рода формулировке существенно придать точный смысл с таким расчетом, чтобы с «бесконечно удаленными», или, как иногда говорят, с «идеальными» точками можно было проводить точные и надежные рассуждения, как с обыкновенными точками на плоскости или в пространстве. Иными словами, мы желали бы, чтобы все правила поведения точек, прямых, плоскостей оставались в силе и для «идеальных» геометрических элементов.

В математическом смысле существование «бесконечно удаленных точек» обеспечено, если отчетливо и без взаимных противоречий установлены математические свойства этих вновь вводимых элементов, т. е. их взаимоотношения с «обыкновенными» точками и между собой. Обыкновенно система геометрических аксиом (например, в евклидовой геометрии) вытекает путем абстракции из наблюдений над физическими объектами: таковы следы прикосновения карандаша к бумаге или мела к доске, натянутые нити, световые лучи, твердые стержни и т. п. Свойства, приписываемые аксиомами математическим точкам и прямым, представляют собой в высшей степени упрощенные и идеализированные описания поведения соответствующих им физических «двойников». Через любые два карандашных пятнышка можно провести не одну, а много карандашных «прямых». Если пятнышки становятся все меньше по диаметру, то все такие «прямые» станут трудно отличимыми одна от другой. Вот что мы, собственно говоря, имеем в виду, высказывая в качестве геометрической аксиомы, что «через любые две точки можно провести одну и только одну прямую»: мы при этом говорим об «абстрактных», чисто умозрительных, геометрических точках и прямых. Геометрические точки и прямые обладают гораздо более простыми свойствами, чем какие бы то ни было физические объекты. Упрощение является существенным условием, позволяющим строить геометрию как

дедуктивную дисциплину.

Как уже было отмечено, обыкновенная геометрия точек и прямых весьма осложнена тем обстоятельством, что две параллельные прямые не имеют точки пересечения. Это побуждает нас сделать дальнейшее упрощение в структуре геометрии, расширяя понятие геометрической точки таким образом, чтобы устранить указанное исключение — совершенно так же, как мы расширяли понятие числа с целью устранения ограничений при вычитании и делении. В геометрии, как и в арифметике, мы озабочены неукоснительно сохранением в расширенной области тех законов, какие регулировали отношения в первоначальной области.

Итак, *мы уславливаемся в том, что к обыкновенным точкам всякой прямой добавляем еще одну, «идеальную», точку и будем считать эту точку принадлежащей одновременно всем прямым, параллельным данной, и никаким другим.* Следствием такого условия является то, что *всякая* пара прямых на плоскости теперь уже пересекается в единственной точке: если прямые не параллельны, то в «обыкновенной» точке; если параллельны, то в им обеим принадлежащей «идеальной» точке. По причинам интуитивного порядка эта идеальная точка на прямой называется *бесконечно удаленной точкой* на этой прямой.

Интуитивное представление о точке, удаляющейся в бесконечность по прямой линии, могло бы навести на мысль, что следует добавить *две* идеальные точки на каждой прямой — по одной для каждого направления. Если мы добавляем только одну, то лишь потому, что мы заинтересованы в сохранении закона: через каждые две точки проходит *одна и только одна прямая.* Если бы прямая содержала две бесконечно удаленные точки вместе со всеми, ей параллельными, то вышло бы, что через две такие «точки» проходит бесконечное множество прямых.

Мы уславливаемся также в том, что к обыкновенным прямым на плоскости добавляем еще одну «идеальную», так называемую «бесконечно удаленную» прямую, содержащую все бесконечно удаленные точки плоскости и никаких других. Мы вынуждены принять именно такое условие, если хотим сохранить первоначальный закон — «через всякие две точки проходит одна прямая» и вновь утвержденный закон — «всякие две прямые пересекаются в одной точке». В самом деле, возьмем две какие-нибудь идеальные точки. Единственная прямая, которая должна проходить через эти точки, не может быть обыкновенной прямой, так как по принятому условию каждая обыкновенная прямая содержит только одну идеальную точку. С другой стороны, эта прямая не может содержать обыкновенных точек, так как через обыкновенную точку и одну из идеальных точек непременно прошла бы обыкновенная прямая. Наконец, рассматриваемая прямая непременно содержит *все* идеальные точки, так как мы хотим, чтобы она имела одну общую точку со всякой обыкновенной прямой. Итак, прямая, о которой идет речь, неизбежно

должна обладать как раз всеми теми свойствами, которыми мы наделили идеальную прямую в нашей плоскости.

Согласно принятым условиям, каждая бесконечно удаленная точка определяется или представляется семейством параллельных прямых, точно так же как иррациональное число определяется последовательностью «вложенных» рациональных отрезков. Такого рода условный способ описывать параллельность с помощью терминов, первоначально предназначенных для интуитивно отличных объектов, единственной своей целью имеет сделать излишним перечисление исключительных случаев; эти последние теперь автоматически покрываются теми же терминами (и оборотами речи), которые первоначально употреблялись для «обыкновенных» случаев.

Резюмируем: наши условия, касающиеся бесконечно удаленных элементов, были выбраны таким образом, чтобы законы, регулирующие отношение инцидентности между обыкновенными точками и прямыми, сохранялись и в расширенной области, чтобы операция нахождения точки пересечения двух прямых, ранее возможная только в случае непараллельности, могла быть выполнена без ограничений. Соображения, которые привели нас к формальному упрощению в отношениях инцидентности, способны показаться несколько абстрактными. Но читатель убедится на следующих страницах, что они будут вполне оправданы результатами.

2. Идеальные элементы и проектирование. Введение бесконечно удаленных точек и бесконечно удаленной прямой на плоскости позволит нам гораздо более удовлетворительным образом рассмотреть проектирование одной плоскости на другую. Пусть плоскость π проектируется на плоскость π' из центра O (рис. 83). Эта проекция устанавливает соответствие между точками и прямыми π и точками и прямыми π'. Каждой точке A на π соответствует единственная точка A' на π' со следующими исключениями: если выходящий из O проектирующий луч *параллелен* плоскости π', то он пересекает π в точке A, которой не соответствует никакая обыкновенная точка плоскости π'. Такие исключительные точки плоскости π расположены на прямой l, которой не соответствует никакая обыкновенная прямая плоскости π'. Но оговаривать эти исключения становится излишним, если мы условимся точке A сопоставлять бесконечно удаленную точку на плоскости π', взятую в направлении прямой OA, а прямой l — сопоставлять бесконечно удаленную прямую в плоскости π'. Аналогично, некоторую бесконечно удаленную точку в плоскости π мы сопоставляем каждой точке B' на такой прямой m' в плоскости π', через которую проходят все лучи, выходящие из O и параллельные плоскости π. Самой прямой m' соответствует бесконечно удаленная прямая плоскости π. Таким образом,

посредством введения в плоскости бесконечно удаленных точек и прямой достигается то, что *проекция одной плоскости на другую устанавливает такое соответствие между точками и прямыми двух плоскостей, которое взаимно однозначно без всяких исключений.* (Так устраняются исключения, упомянутые в сноске на стр. 193.) Далее, легко понять, что из принятых соглашений вытекает следствие: *точка лежит на прямой, если проекция точки лежит на проекции прямой.* Отсюда видно, что все теоремы, относящиеся к коллинеарным точкам, конкуррентным прямым и т. д. и говорящие только о точках, прямых и отношениях инцидентности, инвариантны относительно проектирования в расширенном смысле. Это дает возможность оперировать с бесконечно удаленными точками плоскости π, заменяя их соответствующими получающимися при проектировании обыкновенными точками плоскости π'.

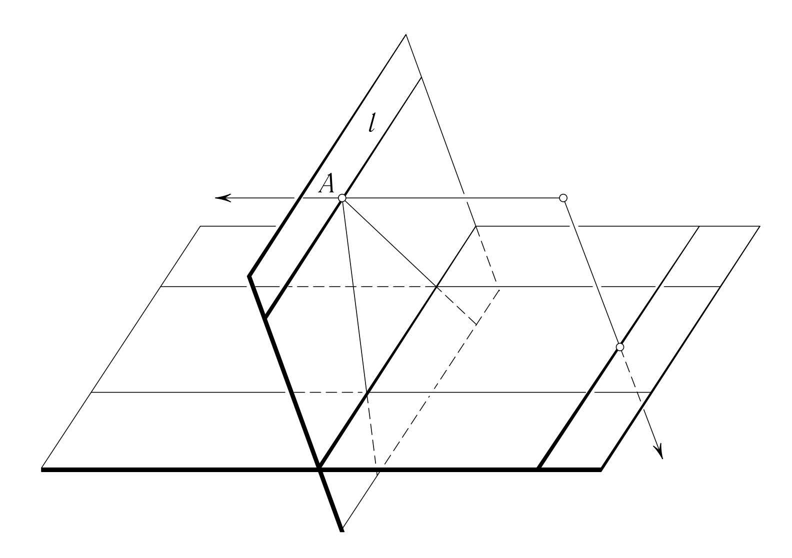


Рис. 83. Возникновение бесконечно удаленных элементов при проектировании

* Можно воспользоваться интерпретацией бесконечно удаленных точек плоскости π с помощью проектирования из внешней точки O на обыкновенные точки другой плоскости π', чтобы получить конкретную евклидову «модель» расширенной плоскости. Для этого не будем обращать внимания на плоскость π', а сосредоточимся на плоскости π и прямых, проходящих через O. Каждой обыкновенной точке π соответствует прямая, проходящая через O, непараллельная π; каждой бесконечно удаленной точке π — прямая, проходящая через O, параллельная π. Итак, совокупности всех точек π, обыкновенных и идеальных, соответствует совокупность прямых, проходящих через O, и это соответ-

ствие взаимно однозначно без всяких исключений. *Точки* на некоторой *прямой* в плоскости π переходят в *прямые* на *плоскости*, проходящей через O. Точка и прямая в плоскости π инцидентны в том и только в том случае, если инцидентны соответствующие прямая и плоскость, проходящие через O. Другими словами, геометрия инцидентности точек и прямых в расширенной плоскости совершенно равносильна геометрии инцидентности обыкновенных прямых и плоскостей, проходящих через фиксированную точку пространства.

Положение вещей в трехмерном пространстве вполне аналогично, хотя отпадает возможность пользоваться наглядным аппаратом проектирования. Здесь тоже мы вводим особую бесконечно удаленную точку, связанную с каждым семейством параллельных прямых. В каждой плоскости имеется бесконечно удаленная прямая. Затем вводится новый элемент — бесконечно удаленная плоскость, состоящая из всех бесконечно удаленных точек пространства и содержащая все бесконечно удаленные прямые. С бесконечно удаленной плоскостью каждая обыкновенная плоскость пересекается по своей собственной бесконечно удаленной прямой.

3. Двойное отношение с бесконечно удаленными элементами. Еще одно замечание следует сделать по поводу двойных отношений с бесконечно удаленными элементами. Будем обозначать символом ∞ бесконечно удаленную точку на прямой l. Посмотрим, как определяется символ $(ABC\infty)$, если A, B, C — три обыкновенные точки на l.

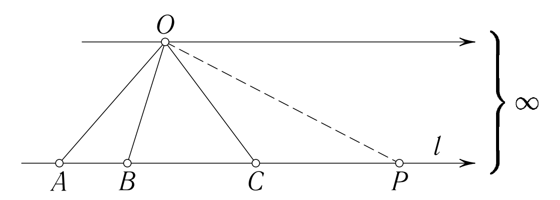


Рис. 84. Двойное отношение с участием бесконечно удаленной точки

Пусть P — некоторая точка на l; тогда $(ABC\infty)$ рассматривается как предел $(ABCP)$, когда P удаляется в бесконечность по l. Но

$$(ABCP) = \frac{CA}{CB} : \frac{PA}{PB},$$

и, когда P неограниченно удаляется, $\frac{PA}{PB}$ стремится к 1. Отсюда вытекает определение:

$$(ABC\infty) = \frac{CA}{CB}.$$

В частности, если $(ABC\infty) = 1$, то C есть середина отрезка AB: *средняя точка отрезка и бесконечно удаленная точка, взятая в направлении отрезка, делят отрезок гармонически.*

Упражнение. Что представляет собой двойное отношение четырех прямых l_1, l_2, l_3, l_4, если они параллельны? Что получится, в частности, с этим двойным отношением, если в качестве l_4 будет взята бесконечно удаленная прямая?

§ 5. Применения

1. Предварительные замечания. После введения бесконечно удаленных элементов уже нет необходимости явно оговаривать все исключительные случаи параллельности, возникающие при построениях и доказательствах теорем. Достаточно помнить, что если точка является бесконечно удаленной, то все проходящие через нее прямые параллельны. Отпадает и необходимость делать различие между центральной и параллельной проекциями, так как параллельная проекция есть не

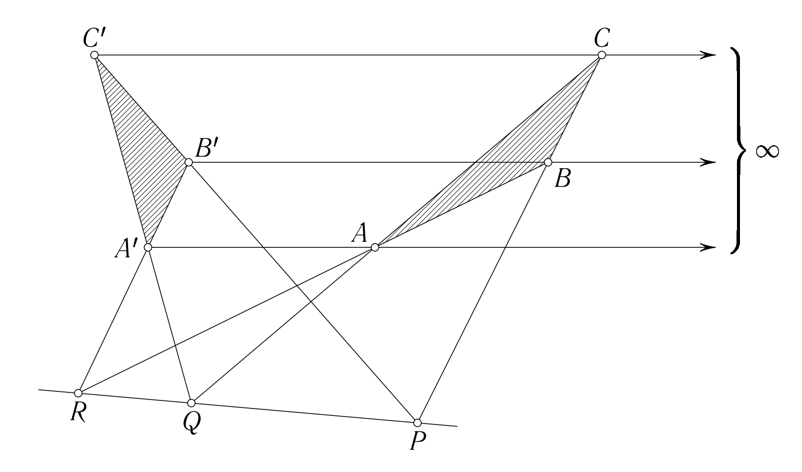


Рис. 85. Дезаргова конфигурация с центром в бесконечности

что иное, как проекция из бесконечно удаленной точки. На рис. 72 точка O или прямая PQR могут оказаться бесконечно удаленными (рис. 85 изображает первый из упомянутых случаев); мы предоставляем читателю в качестве упражнения сформулировать в «финитных» (т. е. не содержащих упоминания о бесконечности) терминах соответствующие утверждения дезарговой теоремы.

Не только формулировка, но и доказательство теоремы, принадлежащей проективной геометрии, нередко упрощаются в результате введения бесконечно удаленных элементов. Общий принцип заключается в следующем. Условимся под «проективным классом» некоторой геометрической фигуры F понимать класс всех фигур, в которые F может быть переведена проективными преобразованиями. Проективные свойства F ничем не отличаются от проективных свойств любой фигуры ее проек-

тивного класса, так как по самому определению проективные свойства сохраняются при проектировании. Таким образом, любая проективная теорема (т. е. теорема, говорящая только о проективных свойствах), которая верна для фигуры F, будет также верна для любого «представителя» проективного класса этой фигуры, и обратно. Поэтому, чтобы доказать такую теорему для F, достаточно доказать ее для некоторого «представителя» проективного класса F. Мы можем воспользоваться указанным обстоятельством и выбрать такого «представителя», для которого доказательство проще, чем для самой фигуры F. Например, произвольные две точки A, B плоскости π могут быть спроектированы в бесконечность из данного центра O, если проектировать на плоскость, параллельную плоскости, проходящей через точки O, A, B; прямые, проходящие через A или через B, при этом превратятся в семейства параллельных прямых. Именно такое предварительное преобразование мы выполним при доказательстве проективных теорем, которыми займемся в этом параграфе.

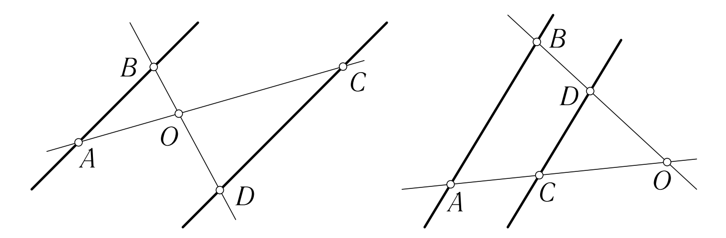


Рис. 86. Подобие треугольников, образованных параллельными прямыми

В дальнейшем нам придется воспользоваться следующим обстоятельством, относящимся к параллельным прямым. Пусть две прямые, проходящие через точку O, пересекаются прямыми l_1 и l_2 в точках A, B, C, D, как показано на рис. 86. Если прямые l_1 и l_2 параллельны, то $\frac{OA}{OC} = \frac{OB}{OD}$; и обратно, если выполнено последнее соотношение, то прямые l_1 и l_2 параллельны. Доказательство, вытекающее из элементарных свойств подобных треугольников, предоставляется читателю.

2. Двумерное доказательство теоремы Дезарга. Докажем теперь, *не прибегая к пространственному проектированию*, что если два треугольника ABC и $A'B'C'$ расположены на плоскости так, как изображено на рис. 72, т. е. если прямые, соединяющие соответствующие вершины, встречаются в одной и той же точке, то точки пересечения соответствующих сторон P, Q, R лежат на одной прямой. Для этого прежде всего спроектируем чертеж таким образом, чтобы точки Q и R ушли в бесконечность. После такого проектирования прямая $A'B'$ станет параллельна прямой AB, а прямая $A'C'$ — прямой AC (рис. 87).

Как было отмечено в пункте 1 настоящего параграфа, чтобы доказать теорему Дезарга в общем случае, достаточно доказать ее только для

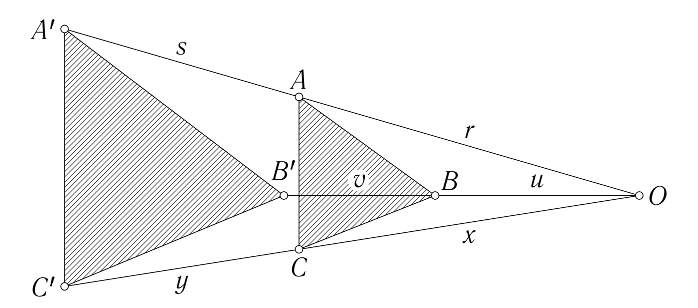


Рис. 87. Доказательство теоремы Дезарга

случая рассматриваемой здесь частной конфигурации. Именно, достаточно показать, что точка P пересечения сторон BC и $B'C'$ также уйдет в бесконечность, т. е. что прямая $B'C'$ параллельна прямой BC: тогда точки P, Q, R будут коллинеарны (так как все три будут лежать на бесконечно удаленной прямой). Обратим внимание на то, что

$$AB \parallel A'B' \quad \text{влечет} \quad \frac{u}{v} = \frac{r}{s},$$

и

$$AC \parallel A'C' \quad \text{влечет} \quad \frac{x}{y} = \frac{r}{s}.$$

Поэтому $\dfrac{u}{v} = \dfrac{x}{y}$, а отсюда следует $BC \parallel B'C'$, что и требовалось доказать.

Следует отметить, что приведенное доказательство теоремы Дезарга опирается на математическое понятие длины отрезка. Таким образом, проективная теорема доказана в данном случае метрическими средствами. Другое заслуживающее внимания обстоятельство заключается в следующем. Мы указывали раньше (стр. 198), что понятию проективного преобразования может быть дано «внутреннее» определение («проективное преобразование плоскости — такое, которое оставляет инвариантными все двойные отношения»): отсюда вытекает, что теорема Дезарга способна быть сформулирована и доказана без выхода в пространство, т. е. без использования трехмерных представлений и построений.

Упражнение. Докажите подобным же образом теорему, обратную дезарговой: если треугольники ABC и $A'B'C'$ таковы, что P, Q, R коллинеарны, то прямые AA', BB', CC' конкуррентны.

3. Теорема Паскаля [1]**.** Эта теорема формулируется так: *если вершины шестиугольника лежат поочередно на двух пересекающихся прямых, то точки P, Q, R пересечения противоположных сторон этого шестиугольника коллинеарны* (рис. 88). (Контур шестиугольника может быть самопересекающимся. Что такое «противоположные» стороны, можно легко понять из схемы на рис. 89.)

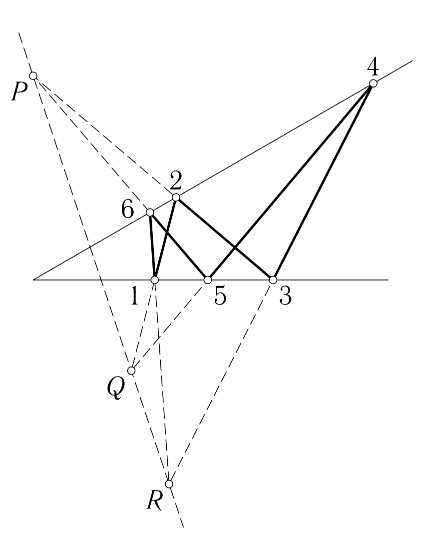


Рис. 88. Конфигурация Паскаля

Выполняя предварительное проектирование, можно допустить, что P и Q ушли в бесконечность. Остается показать, что R также уйдет в бесконечность. Ситуация иллюстрируется рис. 90, где $23 \parallel 56$ и $12 \parallel 45$. Нужно показать, что $16 \parallel 34$. Мы имеем

$$\frac{a}{a+x} = \frac{b+y}{b+y+s},$$
$$\frac{b}{b+y} = \frac{a+x}{a+x+r}.$$

Поэтому

$$\frac{a}{b} = \frac{a+x+r}{b+y+s},$$

так что $16 \parallel 34$, что и требовалось доказать.

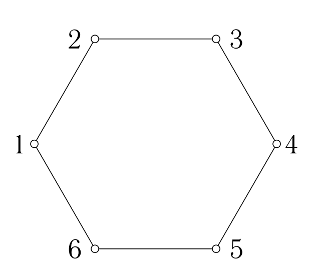


Рис. 89. Нумерация вершин шестиугольника

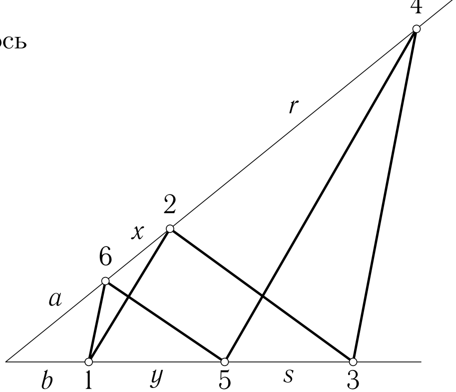


Рис. 90. Доказательство теоремы Паскаля

[1] На стр. 230 будет рассмотрена более общая теорема этого же типа. Настоящий частный случай связывается также с именем Паппа Александрийского (III столетие до нашей эры).

4. Теорема Брианшона. Эта теорема формулируется так: *если стороны шестиугольника проходят поочередно через две данные точки P и Q, то три диагонали, соединяющие противоположные вершины шестиугольника, конкуррентны* (рис. 91).

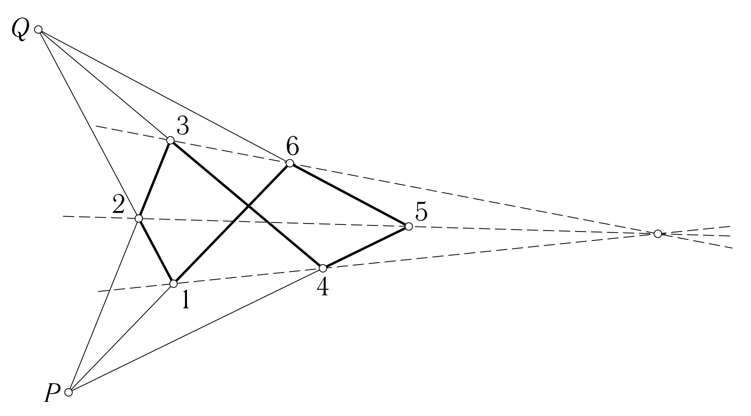


Рис. 91. Конфигурация Брианшона

Посредством предварительного проектирования можно отправить в бесконечность точку P и точку, в которой пересекаются две какие-нибудь диагонали, например 14 и 36. Полученная ситуация изображена на рис. 92. Так как 14 $\parallel$ 36, то $\frac{a}{b} = \frac{u}{v}$. Но вместе с тем $\frac{x}{y} = \frac{a}{b}$ и $\frac{u}{v} = \frac{r}{s}$. Значит, $\frac{x}{y} = \frac{r}{s}$ и поэтому 36 $\parallel$ 25, так что все три диагонали параллельны и, следовательно, конкуррентны. Этого достаточно, чтобы считать теорему доказанной и в общем случае.

5. Замечание по поводу двойственности. Читатель, вероятно, уже заметил замечательное сходство теорем Паскаля (1623–1662) и Брианшона (1785–1864). Это сходство особенно бросается в глаза, если обе формулировки поставить рядом:

Теорема Паскаля	Теорема Брианшона
Если вершины шестиугольника лежат поочередно на двух прямых, то точки пересечения противоположных сторон коллинеарны.	*Если стороны шестиугольника проходят поочередно через две точки, то прямые, соединяющие противоположные вершины, конкуррентны.*

Не только теоремы Паскаля и Брианшона, но все вообще теоремы проективной геометрии группируются попарно таким образом, что две теоремы одной и той же пары сходны между собой и, так сказать, идентичны по своей структуре. Это явление носит название *двойственности*.

В геометрии плоскости точка и прямая представляют собой *взаимно двойственные элементы*. Провести прямую через точку и отметить точку на прямой — операции *взаимно двойственные*. Две фигуры взаимно двойственны, если одна может быть получена из другой посредством замены каждого элемента и каждой операции двойственным элементом и двойственной операцией. Две теоремы взаимно двойственны, если одна превращается в другую при замене каждого элемента и каждой операции двойственным элементом и двойственной операцией. Например, теоремы Паскаля и Брианшона взаимно двойственны, тогда как теоремой, двойственной теореме Дезарга, является теорема, ей обратная. Явление двойственности резко отличает проективную геометрию от элементарной (метрической), в которой никакой двойственности не наблюдается. (Например, было бы бессмысленно искать какое-нибудь «двойственное» утверждение по отношению к тому факту, что данный угол содержит 37° или что данный отрезок равен 2 линейным единицам.) *Принцип двойственности*, согласно которому *каждой верной теореме проективной геометрии сопоставляется двойственная ей, также верная теорема*, во многих учебниках подчеркивается тем, что формулировки взаимно двойственных теорем, вместе со взаимно двойственными их доказательствами, приводятся рядом, как мы это сделали выше. Внутренняя причина явления двойственности будет изучена в следующем параграфе (см. также стр. 228).

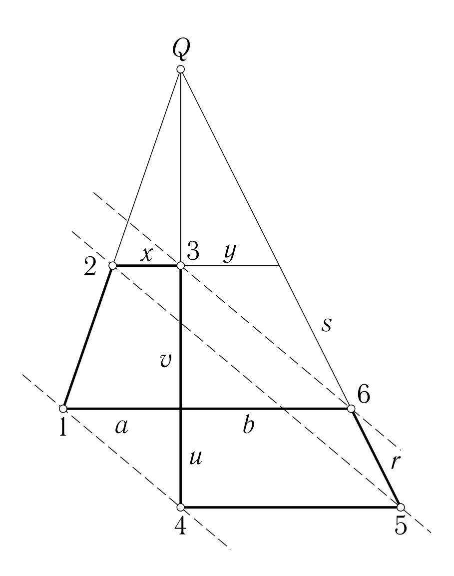


Рис. 92. Доказательство теоремы Брианшона

§ 6. Аналитическое представление

1. Вводные замечания. В раннем периоде развития проективной геометрии существовала настойчиво проводимая тенденция выполнять все построения на синтетической или, как говорилось, «чисто геометрической» основе, вовсе избегая применения чисел и алгебраических методов. Выполнение этой программы встретило на своем пути большие затруднения, так как всегда оставались какие-то пункты, в которых алгебраические формулировки казались неизбежными. Полный успех

в построении чисто синтетической проективной геометрии был достигнут только к концу XIX в. и только ценой значительных осложнений. В этом отношении методы аналитической геометрии оказались гораздо более плодотворными. Для современной математики характерна иная тенденция — положить в основу построения понятие числа, и в геометрии эта тенденция, идущая от Ферма и Декарта, возымела решительный триумф. Аналитическая геометрия перестала быть подсобным аппаратом, играющим служебную роль в геометрических рассуждениях, и стала самостоятельной областью, в которой интуитивная геометрическая интерпретация операций и результатов уже не является последней и окончательной целью, а принимает на себя функцию руководящего принципа, помогающего угадывать и понимать аналитические факты. Такое изменение значения геометрии есть последствие постепенного развития геометрии в историческом плане — развития, широко раздвинувшего рамки классических концепций; оно же обусловило вместе с тем почти органическое слияние геометрии и анализа.

В аналитической геометрии под «координатами» геометрического объекта понимается *какая угодно* совокупность чисел, позволяющая определить этот объект *однозначно*. Так, точка определяется своими прямоугольными координатами x, y или своими полярными координатами ρ, θ; с другой стороны, например, треугольник определяется координатами трех вершин, что в целом составляет шесть координат. Мы знаем, что прямая линия в плоскости x, y представляет собой геометрическое место всех точек $P(x,\, y)$ (об обозначениях см. стр. 92), координаты которых удовлетворяют некоторому линейному уравнению

$$ax + by + c = 0. \tag{1}$$

Поэтому можно три числа a, b, c назвать «координатами» этой прямой. Например, $a = 0$, $b = 1$, $c = 0$ определяют прямую $y = 0$, т. е. ось x; $a = 1$, $b = -1$, $c = 0$ определяют прямую $x = y$, которая делит пополам угол между положительной осью x и положительной осью y. Таким же образом следующие уравнения определяют «конические сечения»: $x^2 + y^2 = r^2$ — окружность радиуса r с центром в начале координат, $(x - a)^2 + (y - b)^2 = r^2$ — окружность радиуса r с центром $(a,\, b)$, $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$ — эллипс и т. д.

Более или менее наивный подход к аналитической геометрии заключается в том, чтобы, отправляясь от чисто «геометрических» представлений — точка, прямая и т. д., — переводить их затем на язык чисел. Современная точка зрения противоположна. Мы отправляемся от *множества всевозможных пар чисел* x, y и *называем* каждую такую пару *точкой*, так как можем, если пожелаем, наглядно *интерпретировать* такую пару чисел с помощью общедоступного понятия геометрической

точки. Точно так же прямая линия является геометрическим представлением или интерпретацией линейного уравнения, связывающего x и y. Указанный перенос акцента от интуитивного понимания геометрии к аналитическому открывает возможность, в частности, простой и вполне строгой трактовки бесконечно удаленных точек в проективной геометрии; он же необходим для более глубокого проникновения в эту область. Для тех читателей, которые обладают достаточной предварительной математической подготовкой, мы дадим теперь некоторый очерк применения аналитических методов в проективной геометрии.

***2. Однородные координаты. Алгебраические основы двойственности.** В обыкновенной аналитической геометрии прямоугольными координатами точки на плоскости являются снабженные знаками расстояния точки от двух взаимно перпендикулярных осей. Но в такой системе координат не находится места для бесконечно удаленных точек расширенной проективной плоскости. Поэтому, если мы хотим пользоваться аналитическими методами в проективной геометрии, то необходимо найти такую координатную систему, которая смогла бы включить идеальные точки наравне с обыкновенными. Легче всего дать описание такой координатной системы, если представить себе данную плоскость X, Y (которую будем обозначать через π) расположенной в трехмерном пространстве с прямоугольными координатами x, y, z (эти буквы обозначают снабженные знаками расстояния точки от трех координатных плоскостей, образованных осями x, y и z). Представим себе, что плоскость π параллельна координатной плоскости x, y и находится на расстоянии 1 от нее, так что трехмерные координаты точки P в плоскости π будут $(X, Y, 1)$. Принимая начало O координатной системы за центр проектирования, заметим, что *всякой точке P взаимно однозначно соответствует некоторая прямая OP, проходящая через начало координат* (см. стр. 98). В частности, бесконечно удаленным точкам плоскости π соответствуют прямые, проходящие через O и параллельные π.

Посмотрим теперь, что же представляет собой система однородных координат для точек плоскости π. Чтобы найти однородные координаты обыкновенной точки P в этой плоскости, возьмем прямую OP и на ней выберем *произвольную* точку Q, отличную от O (рис. 93). Обыкновенные трехмерные координаты x, y, z точки Q считаются *однородными координатами* точки P в плоскости π. В частности, координаты $(X, Y, 1)$ самой точки P являются ее однородными координатами. Но вместе с тем ее же однородными координатами явятся любые числа (tX, tY, t), где $t \neq 0$, так как координаты всех точек прямой OP (кроме O) имеют как раз такой вид. (Мы исключаем точку $(0, 0, 0)$, потому что она лежит на всех прямых, проходящих через O, и не может служить

для их различения.)

Система однородных координат, конечно, представляет известное неудобство в том отношении, что нужны три числа вместо двух для определения точки, и, самое главное, координаты точки определяются не однозначно, а с точностью до постоянного множителя. Но она имеет то безусловное преимущество, что она охватывает и идеальные, бесконечно удаленные точки плоскости π. Действительно, такой идеальной точке P

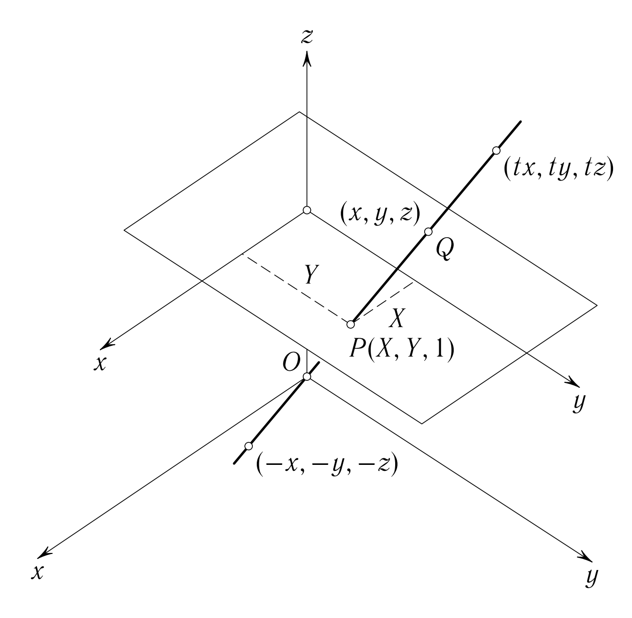


Рис. 93. Однородные координаты

соответствует некоторая прямая, проходящая через O и параллельная π; всякая точка Q на такой прямой имеет координаты вида $(x,\ y,\ 0)$; таким образом, *однородные координаты идеальных точек плоскости* π *имеют вид* $(x,\ y,\ 0)$. Нетрудно написать в однородных координатах уравнение прямой линии на плоскости π. Для этого достаточно заметить, что прямые, соединяющие O с точками этой прямой, лежат в плоскости, проходящей через O. В аналитической геометрии доказывается, что уравнение такой плоскости имеет вид

$$ax + by + cz = 0. \qquad (1')$$

Это же есть и уравнение данной прямой в однородных координатах.

Теперь, когда геометрическая модель, изображающая точки плоскости π в виде прямых, проходящих через O, отслужила свою службу, можно ее отбросить и дать следующее чисто аналитическое определение расширенной плоскости:

Точка есть не что иное, как тройка действительных чисел (x, y, z), из которых не все равны нулю. Две такие тройки (x_1, y_1, z_1) и (x_2, y_2, z_2) определяют *одну и ту же* точку, если существует такое число $t \neq 0$, что

$$x_2 = tx, \quad y_2 = ty_1, \quad z_2 = tz_1.$$

Другими словами, можно, не меняя самой точки, умножать ее координаты на произвольный множитель, отличный от нуля. (Потому эти координаты и называются *однородными.*) Точка (x, y, z) *обыкновенная*, если z отлично от нуля, и *идеальная*, если z равно нулю.

Прямая линия в плоскости π состоит из всех точек (x, y, z), удовлетворяющих линейному уравнению вида

$$ax + by + cz = 0, \tag{1'}$$

где a, b, c — постоянные числа, не все равные нулю. В частности, бесконечно удаленные точки плоскости π удовлетворяют уравнению

$$z = 0; \tag{2}$$

согласно определению, это — также уравнение прямой, именно — *бесконечно удаленной прямой* плоскости π. Так как прямая определяется уравнением вида (1′), то тройка чисел (a, b, c) может быть рассматриваема как *однородные координаты прямой* (1′). Далее следует, что при произвольном $t \neq 0$ тройка чисел (ta, tb, tc) представляет собой координаты той же прямой, так как уравнение

$$(ta)x + (tb)y + (tc)z = 0 \tag{3}$$

удовлетворяется в точности теми же координатными тройками (x, y, z), что и уравнение (1′).

В этих определениях обнаруживается полная симметрия между точкой и прямой: и та и другая определяются тройкой чисел — однородными координатами (u, v, w). Условие того, что точка (x, y, z) лежит на прямой (a, b, c), выражается равенством

$$ax + by + cz = 0,$$

и это же есть вместе с тем условие того, что точка с координатами (a, b, c) лежит на прямой с координатами (x, y, z). Например, арифметическое тождество

$$2 \cdot 3 + 1 \cdot 4 + (-5) \cdot 2 = 0$$

означает, что точка $(3, 4, 2)$ лежит на прямой $(2, 1, -5)$, и в равной мере, что точка $(2, 1, -5)$ лежит на прямой $(3, 4, 2)$. Эта симметрия и представляет собой основу двойственности между точкой и прямой в проективной геометрии, так как всякое соотношение между точками и прямыми становится некоторым соотношением между прямыми и точками, если координаты точек считать координатами прямых, а координаты

прямых — координатами точек. Толкуя по-новому те же алгебраические операции и результаты, мы получаем теоремы, соответствующие первоначальным в смысле двойственности. Необходимо заметить, с другой стороны, что в обыкновенной плоскости X, Y ни о какой двойственности не может быть речи, так как уравнение прямой в обыкновенных координатах

$$aX + bY + c = 0$$

несимметрично относительно X, Y и a, b, c. Только включение в рассмотрение бесконечно удаленных элементов (точек и прямой) обеспечивает применимость принципа двойственности.

Чтобы перейти от однородных координат x, y, z обыкновенной точки P в плоскости π к обыкновенным прямоугольным координатам, мы просто полагаем $X = \frac{x}{z}$, $Y = \frac{y}{z}$. Тогда X, Y обозначают расстояния точки P от двух перпендикулярных осей в плоскости π, параллельной x- и y-осям, как показано на рис. 93. Мы знаем, что уравнение

$$aX + bY + c = 0$$

представляет прямую в плоскости π. Полагая $X = \frac{x}{z}$, $Y = \frac{y}{z}$ и умножая на z, мы найдем, что уравнение той же прямой в однородных координатах будет

$$ax + by + cz = 0,$$

как это уже было указано на стр. 214. Так, уравнение прямой $2x - 3y + z = 0$ в обыкновенных прямоугольных координатах X, Y примет вид $2X - 3Y + 1 = 0$. Разумеется, последнему уравнению бесконечно удаленная точка рассматриваемой прямой с однородными координатами $(3,\ 2,\ 0)$ уже не отвечает.

Остается сказать еще одно. Нам удалось получить чисто аналитическое определение точки и прямой; но что можно сказать о важном понятии проективного преобразования? Можно установить, что проективное преобразование, понимаемое в том смысле, как это было разъяснено на стр. 197, задается аналитически системой линейных уравнений

$$\left.\begin{aligned} x' &= a_1x + b_1y + c_1z \\ y' &= a_2x + b_2y + c_2z \\ z' &= a_3x + b_3y + c_3z \end{aligned}\right\}, \tag{4}$$

связывающих однородные координаты x', y', z' точек в плоскости π' с однородными координатами x, y, z точек в плоскости π. С аналогичной точки зрения можно *определить* проективное преобразование как такое, которое задается системой уравнений вида (4). Теоремы проективной геометрии тогда становятся теоремами, говорящими о поведении числовых троек $(x,\ y,\ z)$ при таких преобразованиях. Например, доказательство инвариантности двойного отношения при проективных преобразованиях превращается в легкое упражнение из области алгебры линейных преобразований. Не будем вникать в детали этой аналитической процедуры и вернемся вместо того назад — к проективной геометрии в ее более наглядном аспекте.

§ 7. Задачи на построение с помощью одной линейки

В следующих построениях предполагается, что единственным инструментом служит линейка.

Задачи 1–18 заимствованы из одной работы Я. Штейнера, в которой он доказывает, что при геометрических построениях можно обойтись без циркуля, если задан фиксированный круг с центром (см. главу III, стр. 173). Читателю рекомендуется проделать эти задачи в указанном порядке.

Четверка прямых a, b, c, d, проходящих через точку P, называется *гармонической*, если двойное отношение $(abcd)$ равно -1. В этом случае говорят, что c, d *гармонически сопряжены* с a, b и обратно.

1) Докажите, что если в гармонической четверке a, b, c, d прямая a делит пополам угол между c и d, то прямая b перпендикулярна к прямой a.

2) Постройте четвертую гармоническую к трем данным прямым, проходящим через одну точку. (*Указание*: воспользуйтесь теоремой о полном четырехстороннике.)

3) Постройте четвертую гармоническую к трем данным точкам на одной прямой.

4) Даны прямой угол и произвольный угол с общей вершиной и одной общей стороной. Удвойте данный произвольный угол.

5) Дан угол и его биссектриса b. Постройте перпендикуляр к b в вершине данного угла.

6) Докажите, что если проходящие через точку P прямые l_1, l_2, ..., l_n пересекают прямую a в точках A_1, A_2, ..., A_n и прямую b в точках B_1, B_2, ..., B_n, то все точки пересечения пар прямых A_iB_k и A_kB_i ($i \neq k$; $k = 1, 2, \ldots, n$) лежат на одной прямой.

7) Докажите, что если в треугольнике ABC прямая, параллельная стороне BC, пересекает AB в точке B' и AC в точке C', то прямая, соединяющая точку A с точкой D пересечения прямых $B'C$ и $C'B$, делит пополам BC.

7а) Сформулируйте и докажите теорему, обратную 7.

8) На прямой l даны три такие точки P, Q, R, что Q есть середина отрезка PR. Постройте прямую, параллельную l и проходящую через данную точку S.

9) Даны две параллельные прямые l_1 и l_2; разделите пополам данный отрезок AB на прямой l_1.

10) Через данную точку P провести прямую, параллельную двум данным параллельным между собой прямым l_1 и l_2. (*Указание*: используйте 7.)

11) Штейнер предлагает следующее решение задачи об удвоении данного отрезка AB при условии, что задана прямая l, параллельная AB: через точку C, не лежащую ни на прямой l, ни на прямой AB, провести прямые CA и CB; пусть A_1 и B_1 — соответственно точки их пересечения с прямой l. Затем (см. 10) провести через C прямую, параллельную l; пусть D — точка ее пересечения с BA_1. Если E — точка пересечения AB и DB_1, то $AE = 2 \cdot AB$.

Докажите последнее утверждение.

12) Разделите отрезок AB на n равных частей, если задана прямая l, параллельная AB. (*Указание*: пользуясь 11, отложите сначала n раз данный

отрезок на прямой l.)

13) Дан параллелограмм $ABCD$. Через данную точку P проведите прямую, параллельную данной прямой l. (*Указание*: примените 10 к центру параллелограмма и воспользуйтесь 8.)

14) Дан параллелограмм; увеличьте данный отрезок в n раз. (*Указание*: примените 13 и 11.)

15) Дан параллелограмм; разделите данный отрезок на n равных частей.

16) Дан неподвижный круг с центром. Проведите через данную точку прямую, параллельную данной прямой. (*Указание*: примените 13.)

17) Дан неподвижный круг с центром. Увеличьте и уменьшите данный отрезок в n раз. (*Указание*: примените 13.)

18) Дан неподвижный круг с центром. Проведите через данную точку перпендикуляр к данной прямой. (*Указание*: воспользуйтесь прямоугольником, вписанным в данный круг, с двумя сторонами, параллельными данной прямой, и сведите к предшествующим задачам.)

19) Пересмотрев задачи 1–18, перечислите, какие основные задачи на построение можно выполнить с помощью двусторонней линейки (с двумя параллельными сторонами).

20) Две данные прямые l_1 и l_2 пересекаются в точке P, находящейся *за пределами* чертежа. Постройте прямую, соединяющую данную точку Q с точкой P. (*Указание*: дополните заданные элементы таким образом, чтобы получилась конфигурация плоскостной теоремы Дезарга, причем P и Q стали бы точками пересечения взаимно соответствующих сторон двух треугольников.)

21) Проведите прямую через две точки, между которыми расстояние больше, чем длина линейки. (*Указание*: примените 20.)

22) Прямые l_1 и l_2 пересекаются в точке P; прямые m_1 и m_2 — в точке Q; обе точки P и Q — за пределами чертежа. Постройте ту часть прямой PQ, которая находится в пределах чертежа. (*Указание*: чтобы получить точку прямой PQ, постройте конфигурацию Дезарга таким образом, чтобы две стороны одного треугольника лежали соответственно на l_1 и m_1, две стороны другого — соответственно на l_2 и m_2).

23) Решите 20 с помощью теоремы Паскаля (стр. 209). (*Указание*: достройте конфигурацию Паскаля, рассматривая l_1 и l_2 как пару противоположных сторон шестиугольника, а Q — как точку пересечения другой пары противоположных сторон.)

*24) Каждая из двух прямых, целиком лежащих за пределами чертежа, задана двумя парами прямых линий, пересекающихся за пределами чертежа в точках соответствующей прямой. Определите точку их пересечения с помощью двух прямых, пересекающихся за пределами чертежа.

§ 8. Конические сечения и квадрики

1. Элементарная метрическая геометрия конических сечений. До сих пор мы занимались только точками, прямыми, плоскостями и фигурами, составленными из конечного числа этих элементов. Если бы проективная геометрия ограничивалась рассмотрением таких «ли-

нейных» фигур, она была бы сравнительно малоинтересна. Но фактом первостепенного значения является то обстоятельство, что проективная геометрия этим не ограничивается, а включает также обширную область конических сечений и их многомерных обобщений. Аполлониева метрическая трактовка конических сечений — эллипсов, гипербол и парабол — была одним из выдающихся успехов античной математики. Едва ли можно переоценить значение конических сечений как для чистой, так и для прикладной математики (например, орбиты планет и орбиты электронов в атоме водорода являются коническими сечениями). Не приходится удивляться тому, что классическая, возникшая в Древней Греции, теория конических сечений и в наши дни составляет необходимую часть математического образования. Но греческая геометрия никоим образом не сказала последнего слова. Через две тысячи лет были открыты замечательные проективные свойства конических сечений. Несмотря на простоту и изящество этих свойств, академическая инерция до настоящего времени служит препятствием их проникновению в школьное преподавание.

Начнем с того, что напомним метрические определения конических течений. Таких определений несколько, и их эквивалентность доказывается в элементарной геометрии. Наиболее распространенные определения связаны с *фокусами* кривых. *Эллипс* определяется как геометрическое место таких точек P на плоскости, что сумма их расстояний r_1 и r_2 от двух данных точек F_1 и F_2, называемых фокусами, имеет постоянное значение. (Если фокусы совпадают, кривая превращается в окружность.) *Гипербола* определяется как геометрическое место таких точек P на плоскости, что абсолютная величина разности $r_1 - r_2$ равно одной и той же постоянной величине. *Парабола* определяется как геометрическое место точек P, расстояние которых r от данной точки F равно расстоянию от данной прямой l.

В аналитической геометрии эти кривые представляются уравнениями второй степени относительно прямоугольных координат x, y. Нетрудно доказать, обратно, что всякая кривая, представляемая уравнением второго порядка

$$ax^2 + by^2 + cxy + dx + ey + f = 0,$$

есть или одно из трех названных выше конических сечений, или прямая линия, или пара прямых, или сводится к одной точке, или носит чисто мнимый характер. Как показывается во всяком курсе аналитической геометрии, для доказательства достаточно сделать надлежащим образом подобранную замену координатной системы.

Указанные выше определения конических сечений — существенно метрические, так как пользуются понятием расстояния. Но вот другое определение, устанавливающее место конических сечений в проективной

геометрии: *конические сечения суть не что иное, как проекции окружности на плоскость.* Если мы станем проектировать окружность C из некоторой точки O, то проектирующие прямые образуют бесконечный двойной конус, и пересечение этого конуса с плоскостью π будет проекцией окружности C. Кривая пересечения будет эллипсом или гиперболой, смотря по тому, пересечет ли плоскость только одну «полость» конуса или обе. Возможен и промежуточный случай параболы, если плоскость π параллельна одной из проектирующих прямых, проведенных через O (рис. 94).

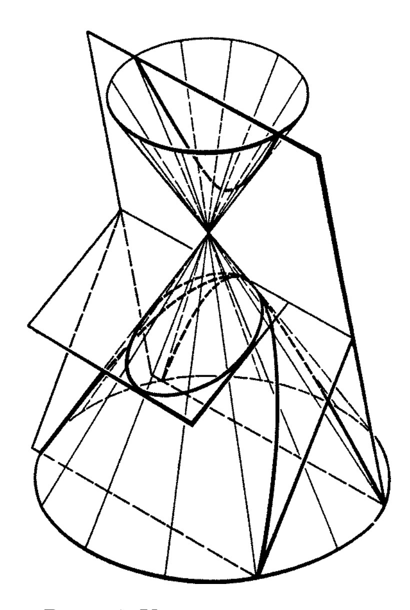


Рис. 94. Конические сечения

Проектирующий конус не обязан быть «прямым круговым» с вершиной O, расположенной вертикально над центром окружности C: он может быть и «наклонным». Но во всех случаях (как мы примем здесь, не приводя доказательства) в пересечении конуса с плоскостью получается кривая, уравнение которой — второй степени; и обратно, всякая кривая второго порядка может быть получена из окружности посредством проектирования. По этой именно причине кривые второго порядка иначе называются коническими сечениями.

Мы уже отметили, что если плоскость пересекает только одну «полость» прямого кругового конуса, то пересечение E представляет собой эллипс. Нетрудно установить, что кривая E удовлетворяет обыкновенному фокальному определению эллипса, которое было сформулировано выше. Приведем очень простое и изящное доказательство, данное в 1822 г. бельгийским математиком Данделеном. Представим себе две сферы S_1 и S_2 (рис. 95), которые касаются плоскости сечения π соответственно в точках F_1 и F_2 и, кроме того, касаются конуса вдоль параллельных окружностей K_1 и K_2. Взяв произвольную точку P кривой E, проведем отрезки PF_1 и PF_2. Затем рассмотрим отрезок PO, соединяющий точку P с вершиной конуса O; этот отрезок целиком лежит на поверхности конуса; обозначим через Q_1 и Q_2 точки его пересечения с окружностями K_1 и K_2. Так как PF_1 и PQ_1 — две

касательные, проведенные из точки P к одной и той же сфере S_1, то

$$PF_1 = PQ_1.$$

Точно так же

$$PF_2 = PQ_2.$$

Складывая эти равенства, мы получаем:

$$PF_1 + PF_2 = PQ_1 + PQ_2.$$

Но $PQ_1 + PQ_2 = Q_1Q_2$ есть расстояние между параллельными окружностями K_1 и K_2 на поверхности конуса: оно не зависит от выбора точки P на кривой E. Отсюда следует, что, какова бы ни была точка P на E, имеет

место равенство

$$PF_1 + PF_2 = \text{const},$$

а это и есть фокальное определение эллипса. Итак, E есть эллипс, а F_1 и F_2 — его фокусы.

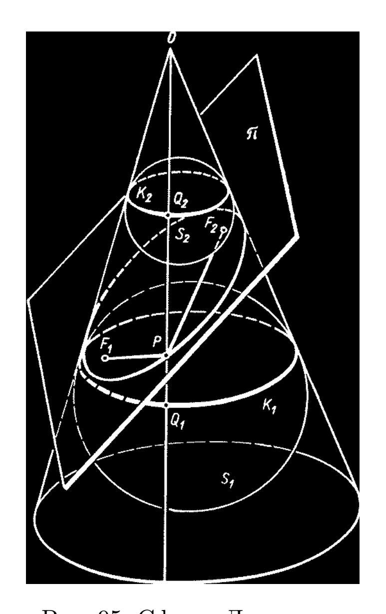


Рис. 95. Сферы Данделена

Упражнение. Если плоскость пересекает обе «полости» конуса, то кривая пересечения — гипербола. Докажите это утверждение, помещая по одной сфере в каждой из «полостей» конуса.

2. Проективные свойства конических сечений. Основываясь на положениях, установленных в предыдущем пункте, примем теперь временно следующее определение: коническое сечение есть проекция окружности на плоскость. Это определение в большей степени отвечает духу проективной геометрии, чем общепринятые фокальные определения, так как эти последние всецело опираются на метрическое понятие расстояния. Новое определение тоже не вполне свободно от этого недостатка, поскольку «окружность» — также метрическое понятие. Но через мгновение мы придем к чисто проективному определению конических сечений.

Раз мы приняли, что коническое сечение есть не что иное, как проекция окружности (другими словами, под термином «коническое сечение» мы понимаем любую кривую, принадлежащую проективному

классу окружности; см. стр. 206), то отсюда сейчас же следует, что всякое свойство окружности, инвариантное относительно проективных преобразований, должно также принадлежать любому коническому сечению. Вспомним теперь следующее хорошо известное — метрическое — свойство окружности: «вписанные в окружность углы, опирающиеся на одну и ту же дугу, равны между собой». На рис. 96 угол AOB, опирающийся на дугу AB, не зависит от положения точки O на окружности. Свяжем, дальше, указанное обстоятельство с проективным понятием двойного отношения, вводя на окружности уже не две точки A, B, а четыре: A, B, C, D. Четыре прямые a, b, c, d, соединяющие эти точки с точкой O на окружности, имеют двойное отношение (a, b, c, d), зависящее только от углов, опирающихся на дуги CA, CB, DA, DB. Соединяя A, B, C, D с какой-нибудь другой точкой O' на окружности, получим прямые a', b', c', d'. Из отмеченного ранее свойства окружности вытекает, что две четверки прямых «конгруэнтны»[1]. Поэтому у них будет одно и то же двойное отношение: $(a'b'c'd') = (abcd)$. Спроектируем окружность на некоторое коническое сечение K: тогда на K получится четверка точек, которые мы снова обозначим через A, B, C, D, две точки O и O' и две четверки прямых a, b, c, d и a', b', c', d'. Эти две четверки прямых уже не будут конгруэнтны, так как углы при проектировании, вообще говоря, не сохраняются. Но так как двойное отношение при проектировании не изменяется, то равенство $(abcd) = (a'b'c'd')$ по-прежнему имеет место. Мы пришли, таким образом, к следующей основной теореме: *если четыре точки конического сечения K, например A, B, C, D, соединены с пятой точкой O того же сечения прямыми a, b, c, d, то двойное отношение $(abcd)$ не зависит от положения O на кривой K* (рис. 97).

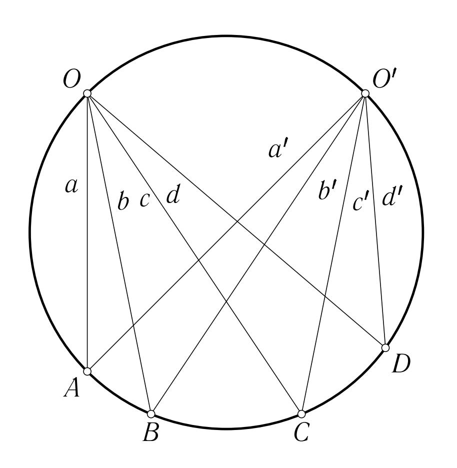


Рис. 96. Двойное отношение на окружности

Это — замечательный результат. Как нам уже известно, если четыре точки A, B, C, D взяты на прямой, то двойное отношение, составленное из соединяющих эти точки с пятой точкой O прямых, не зависит от

[1] Четверка прямых a, b, c, d считается конгруэнтной другой четверке a', b', c', d', если углы между каждой парой прямых в первой четверке равны как по величине, так и по направлению отсчета углам между соответствующими прямыми второй четверки.

выбора этой пятой точки. Это — исходное положение, лежащее в основе проективной геометрии. Теперь мы узнали, что аналогичное утверждение справедливо и относительно четырех точек, взятых на некотором коническом сечении K, однако с существенным ограничением: пятая точка O уже не может свободно двигаться по всей плоскости, а может только перемещаться по коническому сечению K.

Не представляет особого труда доказать и обратную теорему в следующей форме: *если на кривой K имеются две точки O и O', обладающие тем свойством, что какова бы ни была четверка точек A, B, C, D на кривой K, двойные отношения, составленные из прямых, соединяющих эти точки с O, и из прямых, соединяющих эти точки с O', равны между собой, то кривая K есть коническое сечение* (а уж тогда, по прямой теореме, двойное отношение, составленное из прямых, соединяющих четыре данные точки с произвольной точкой O'' на K, будет иметь одно и то же постоянное значение). Но доказательства мы здесь приводить не будем.

Изложенные проективные свойства конических сечений наводят на мысль об общем методе точечного построения этих кривых. Условимся под *пучком прямых* понимать совокупность всех прямых плоскости, проходящих через данную точку O. Рассмотрим пучки прямых, проходящих через две точки O и O', расположенные на коническом сечении K. Между прямыми пучка O и прямыми пучка O' можно установить взаимно однозначное соответствие, сопоставляя прямой a из первого пучка прямую a' из второго всякий раз, как a и a' встречаются в некоторой точке A кривой K. Тогда любая четверка прямых a, b, c, d из пучка O будет иметь то же двойное отношение, что и соответствующая четверка a', b', c', d' из пучка O'. Всякое взаимно однозначное соответствие между двумя пучками прямых, обладающее этим последним свойством, называется *проективным соответствием.* (Это определение двойственно по отношению к определению проективного соответствия между точками на двух прямых, см. стр. 198–198.) Пользуясь этим определением, можно теперь утверждать: *коническое сечение K есть геометрическое место точек пересечения взаимно соответствующих прямых из двух пучков, находящихся в проективном соответствии.* Полученная теорема подводит фундамент под следующее чисто проективное определение конических сечений: *коническим*

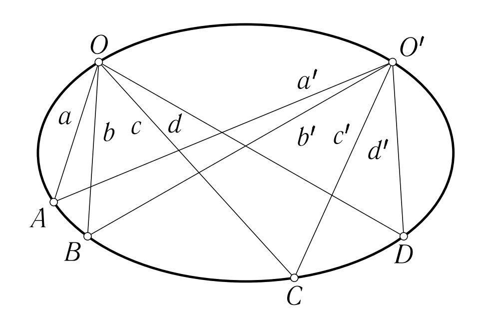


Рис. 97. Двойное отношение на эллипсе

сечением называется геометрическое место точек пересечения взаимно соответствующих прямых из двух пучков, находящихся в проективном соответствии[1]. Как ни соблазнительно проникнуть в глубь теории конических сечений, строящейся на таком определении, однако мы вынуждены ограничиться немногими замечаниями по этому поводу.

Пары пучков, находящихся в проективном соответствии, можно получить следующим образом. Спроектируем все точки P прямой линии l из двух разных центров O и O'' и установим между проектирующими пучками взаимно однозначное соответствие, сопоставляя друг другу те прямые, которые пересекаются на прямой l. Этого достаточно, чтобы полученные пучки находились в проективном соответствии. Затем возьмем пучок O'' и перенесем его «как нечто твердое» в произвольное положение O'. Что новый пучок O' будет находиться в проективном соответствии с пучком O, это совершенно очевидно. Но замечательно то, что любое проективное соответствие между двумя пучками можно

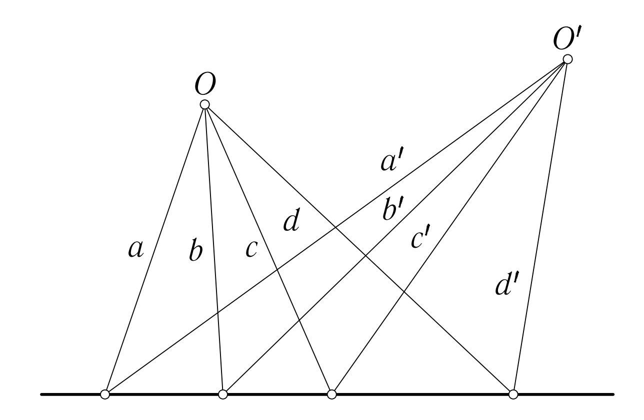


Рис. 98. К построению проективных пучков прямых

получить именно таким образом. (Это обстоятельство двойственно по отношению к упражнению 1 на стр. 199.) Если пучки O и O' конгруэнтны, получается окружность. Если углы между соответствующими лучами в двух пучках равны, но отсчитываются в противоположных направлениях, то получается равносторонняя гипербола (рис. 99).

Следует еще заметить, что указанное определение конического сечения может, в частности, дать и прямую линию, как это показано на рис. 98. В этом случае прямая OO'' соответствует сама себе, и все ее точки должны быть рассматриваемы как принадлежащие искомому геометрическому месту. Таким образом, коническое сечение вырождается в

[1] Это геометрическое место, при известных обстоятельствах, может вырождаться в прямую линию; см. рис. 98.

пару прямых: это обстоятельство вполне согласуется с тем фактом, что существуют сечения конуса, состоящие из двух прямых (если секущая плоскость проходит через вершину конуса).

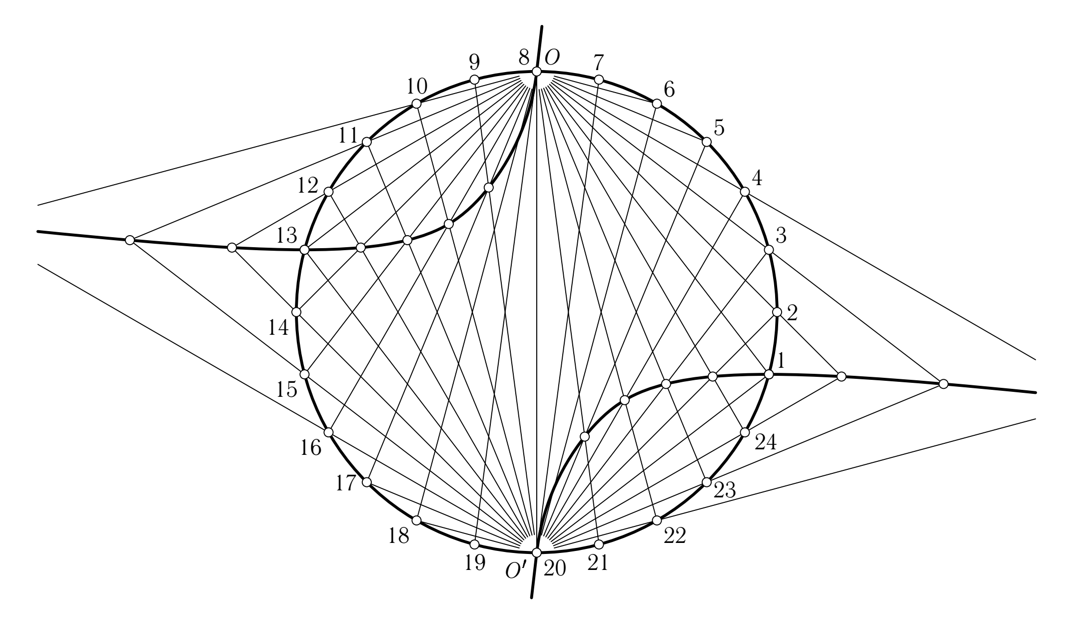


Рис. 99. Образование окружности и равносторонней гиперболы с помощью проективных пучков

Упражнения. 1) Вычертите эллипсы, гиперболы и параболы с помощью проективных пучков. (Читателю настойчиво рекомендуется экспериментировать с подобного рода построениями. Это в высшей степени способствует пониманию сути дела.)

2) Дано пять точек O, O', A, B, C некоторого конического сечения K. Найдите точки пересечения D произвольной прямой d пучка O с кривой K. (*Указание*: через O проведите прямые OA, OB, OC и назовите их a, b, c. Через O' проведите прямые $O'A$, $O'B$, $O'C$ и назовите их a', b', c'. Проведите через O прямую d и постройте такую прямую d' пучка O', что $(abcd) = (a'b'c'd')$. Тогда точка пересечения d и d' принадлежит кривой K.)

3. Конические сечения как «линейчатые кривые». Понятие касательной к коническому сечению принадлежит проективной геометрии, так как касательная к коническому сечению есть прямая, имеющая с самой кривой только одну общую точку, а это — свойство, сохраняющееся при проектировании. Проективные свойства касательных к коническим сечениям основываются на следующей теореме:

Двойное отношение точек пересечения четырех фиксированных касательных к коническому сечению с произвольной пятой касательной

не зависит от выбора этой пятой касательной.

Доказательство этой теоремы весьма просто. Так как любое коническое сечение есть проекция окружности и так как в теореме идет речь только о таких свойствах, которые инвариантны относительно проектирования, то, чтобы доказать теорему в общем случае, достаточно доказать ее для частного случая окружности.

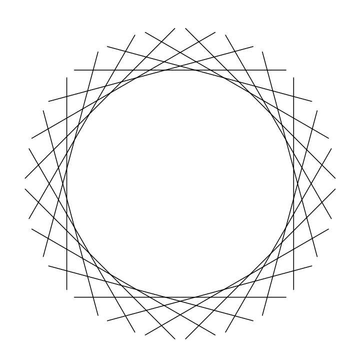


Рис. 100. Окружность как совокупность касательных

Для этого же частного случая теорема доказывается средствами элементарной геометрии. Пусть P, Q, R, S — четыре точки на окружности K; a, b, c, d — касательные в этих точках; T — еще какая-нибудь точка на окружности, o — касательная в ней; пусть, далее, A, B, C, D — точки пересечения касательной o с касательными a, b, c, d. Если M — центр окружности, то, очевидно, $\angle TMA = \frac{1}{2}\angle TMP$, и последнее выражение представляет угол, вписанный в K, опирающийся на дугу TP. Таким же образом $\angle TMB$ представляет угол, вписанный в K и опирающийся на дугу TQ. Следовательно,

$$\angle AMB = \frac{1}{2} \smile PQ,$$

где $\frac{1}{2} \smile PQ$ обозначает угол, вписанный в K и опирающийся на дугу PQ. Отсюда видно, что A, B, C, D проектируются из M четырьмя прямыми, углы между которыми имеют величины, зависящие только от положения точек P, Q, R, S. Но тогда двойное отношение $(ABCD)$ зависит только от четырех касательных a, b, c, d, но не от касательной o. Как раз это и нужно было установить.

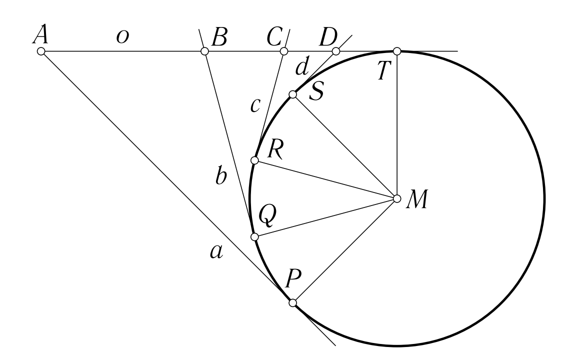


Рис. 101. Свойство касательной к окружности

В предыдущем пункте мы имели случай убедиться, что коническое сечение может быть построено «по точкам», если станем отмечать точки пересечения взаимно соответствующих прямых двух пучков, между которыми установлено проективное соответствие. Только что доказанная теорема дает нам возможность сформулировать двойственную теорему. Возьмем две касательные a и a' к коническому сечению K. Третья касательная t пусть пересекает a и a' соответственно в точках A и A'. Если t будет перемещаться вдоль кривой, то установится соответствие

$$A \longleftrightarrow A'$$

между точками a и точками a'. Это соответствие будет проективным, так как по доказанной теореме произвольная четверка точек на a будет непременно иметь то же двойное отношение, что и соответствующая четверка точек на a'. Отсюда следует, что *коническое сечение* K, рас-

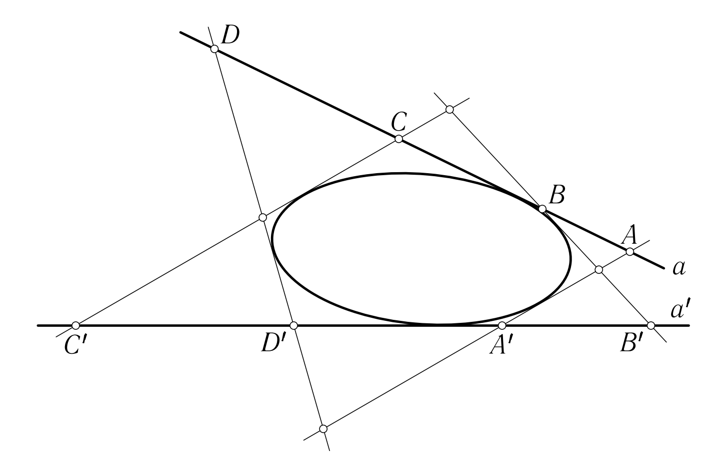


Рис. 102. Проективные ряды точек на двух касательных к эллипсу

сматриваемое как «совокупность своих касательных», «состоит» *из прямых, соединяющих взаимно соответствующие точки двух точечных рядов*[1] *на a и на a'*, находящихся в проективном соответствии. Указанное обстоятельство позволяет ввести новое определение конических сечений, рассматриваемых на этот раз как «линейчатые кривые». Сравним это определение с прежним проективным определением конического сече-

[1] Совокупность точек на прямой называется точечным рядом. Это понятие двойственно по отношению к пучку прямых.

ния, данным в предыдущем пункте:

I	II
Коническое сечение, рассматриваемое как совокупность *точек*, состоит из *точек пересечения взаимно соответствующих прямых* в двух проективных *пучках.*	Коническое сечение, рассматриваемое как «совокупность *прямых*», состоит из *прямых*, *соединяющих взаимно соответствующие точки* в двух проективных *рядах.*

Если мы станем считать касательную к коническому сечению в некоторой его точке двойственным элементом по отношению к самой точке и условимся, кроме того, «линейчатую кривую» (образованную совокупностью касательных) на основе двойственности сопоставлять «точечной кривой» (образованной совокупностью точек), то предыдущие формулировки будут безупречны с точки зрения принципа двойственности. При «переводе» одной формулировки в другую с заменой всех понятий соответствующими двойственными понятиями, «коническое сечение» остается неизменным; но в одном случае оно мыслится как «точечная кривая», определяемая своими точками, в другом — как «линейчатая кривая», определяемая своими касательными.

Из предыдущего вытекает важное следствие: принцип двойственности, первоначально установленный в проективной геометрии плоскости только для точек и прямых, оказывается, может быть распространен и на конические сечения. *Если в формулировке любой теоремы, касающейся точек, прямых и конических сечений, заменить каждый элемент ему двойственным* (не упуская из виду, что точке конического сечения должна быть сопоставляема касательная к этому коническому сечению), *то в результате также получится справедливая теорема.* Пример действия этого принципа мы встретим в пункте 4 настоящего параграфа.

Построение конических сечений, понимаемых как «линейчатые кривые», показано на рис. 103–104. В частности, если в двух проективных точечных рядах бесконечно удаленные точки соответствуют взаимно одна другой (так будет непременно, если точечные ряды конгруэнтны или подобны[1]), то коническое сечение будет параболой; справедливо и обратное утверждение.

Упражнение. Докажите обратную теорему: на двух неподвижных касательных к параболе движущаяся касательная к параболе определяет два подобных точечных ряда.

4. Теоремы Паскаля и Брианшона для общего случая произвольных конических сечений. Одной из лучших иллюстраций

[1] Что такое «конгруэнтные» и «подобные» точечные ряды, достаточно понятно без объяснений.

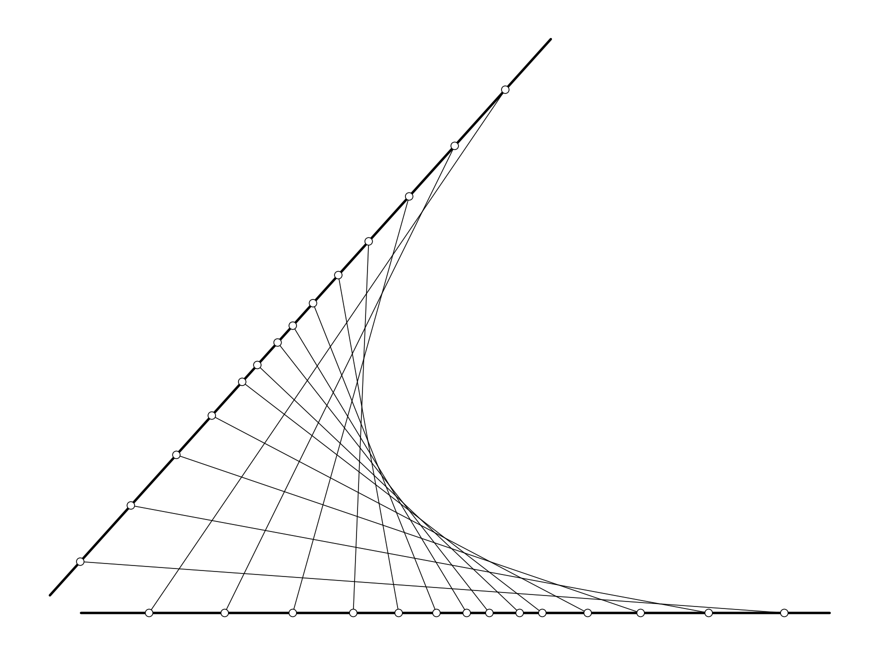


Рис. 103. Парабола, определенная конгруэнтными точечными рядами

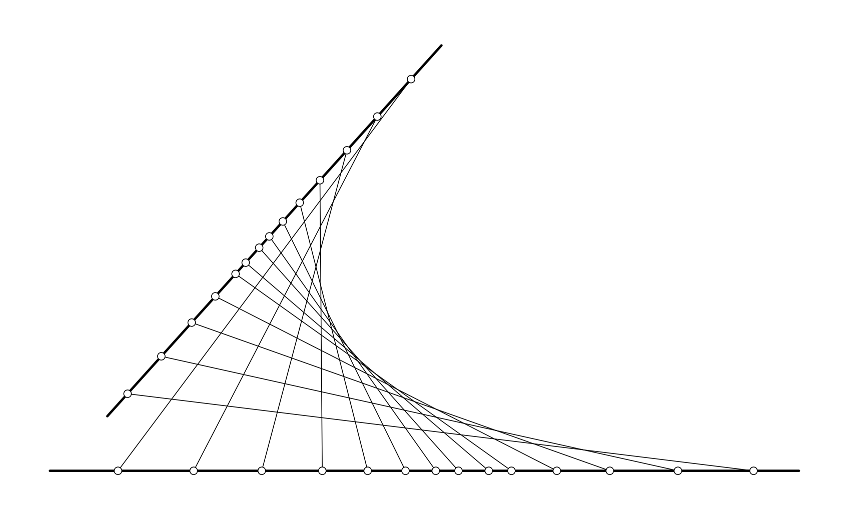


Рис. 104. Парабола, определенная подобными точечными рядами

принципа двойственности применительно к коническим сечениям является взаимоотношение между общими теоремами Паскаля и Брианшона. Первая из них была открыта в 1640 г., вторая — в 1806 г. И, однако, каждая из них есть непосредственное следствие другой, так как всякая теорема, формулировка которой упоминает только конические сечения, прямые и точки, непременно остается справедливой при изменении формулировки по принципу двойственности.

Теоремы, доказанные в § 5 под теми же наименованиями, представляют собой «случаи вырождения» следующих более общих теорем.

Теорема Паскаля. *Противоположные стороны шестиугольника, вписанного в коническое сечение, пересекаются в трех коллинеарных точках.*

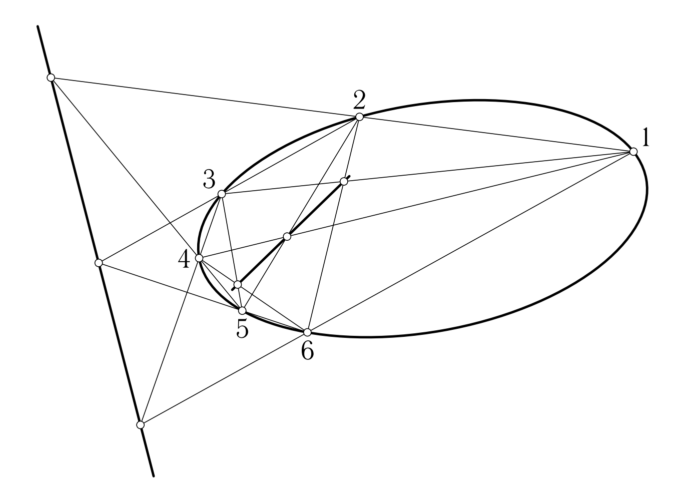


Рис. 105. Общая конфигурация Паскаля. Показаны два случая: один для шестиугольника 1, 2, 3, 4, 5, 6, другой для шестиугольника 1, 3, 5, 2, 6, 4

Теорема Брианшона. *Три диагонали, соединяющее противоположные вершины шестиугольника, описанного около конического сечения, конкуррентны.*

Обе теоремы имеют очевидное проективное содержание. Их двойственность бросается в глаза, если сформулировать их следующим образом:

Теорема Паскаля. Дано шесть точек 1, 2, 3, 4, 5, 6 на коническом сечении. Соединим последовательные точки прямыми (1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 1). Отметим точки пересечения прямых (1, 2) и (4, 5), (2, 3) и (5, 6), (3, 4) и (6, 1). Эти три точки лежат на одной прямой.

Теорема Брианшона. Дано шесть касательных 1, 2, 3, 4, 5, 6 к коническому сечению. Последовательные касательные пересекаются в точках (1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 1). Проведем прямые, соединяющие точки (1, 2) и (4, 5), (2, 3) и (5, 6), (3, 4) и (6, 1). Эти три прямые проходят через одну точку.

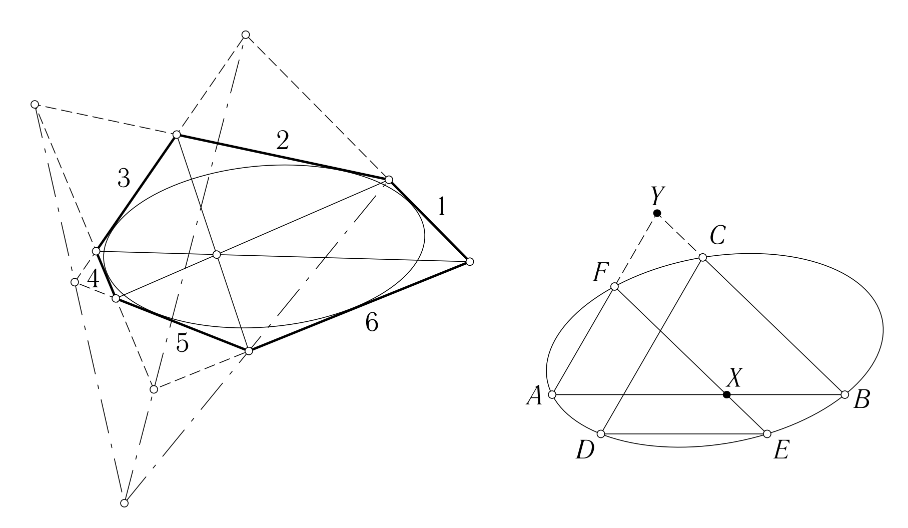


Рис. 106. Общая конфигурация Брианшона. Показаны только два случая

Рис. 107. Доказательство теоремы Паскаля

Доказательства проводятся с помощью специализации такого же рода, как и в рассмотренных раньше случаях вырождения. Докажем теорему Паскаля. Пусть A, B, C, D, E, F — вершины шестиугольника, вписанного в коническое сечение K. Посредством проектирования можно сделать параллельными прямые AB и ED, FA и CD (и тогда получится конфигурация, изображенная на рис. 107; ради удобства шестиугольник на чертеже взят самопересекающимся, хотя в этом нет никакой необходимости.) Нам нужно теперь доказать только одно: что прямая CB параллельна прямой FE; другими словами, что противоположные стороны пересекаются на бесконечно удаленной прямой. Для доказательства рассмотрим четверку точек F, A, B, D, которая, как мы знаем, при проектировании из любой точки K сохраняет одно и то же двойное отношение, скажем, k. Станем проектировать из точки C на прямую AF; получим четверку точек F, A, Y, ∞, причем

$$k = (F,\, A,\, Y,\, \infty) = \frac{YF}{YA}$$

(см. стр. 205).

Станем теперь проектировать из точки E на прямую BA; получим

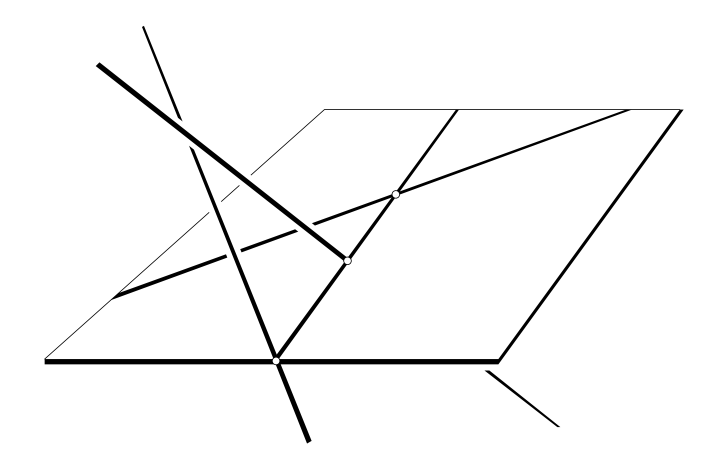


Рис. 108. Построение прямых, пересекающих три данные прямые общего положения

четверку точек X, A, B, ∞, причем

$$k = (X,\ A,\ B,\ \infty) = \frac{BX}{BA}.$$

Итак,

$$\frac{BX}{BA} = \frac{YF}{YA},$$

что как раз и означает, что $YB \parallel FX$. Доказательство теоремы Паскаля закончено.

Теорема Брианшона, как было указано, следует из теоремы Паскаля по принципу двойственности. Но ее можно доказать и непосредственно — путем рассуждения, двойственного относительно только что приведенного. Провести это рассуждение во всех деталях будет прекрасным упражнением для читателя.

5. Гиперболоид. В трехмерном пространстве мы встречаемся с так называемыми *квадриками* (поверхностями второго порядка), которые в данном случае играют ту же роль, что «конические сечения» (кривые второго порядка) на плоскости.

Простейшими из них являются сфера и эллипсоид. Квадрики более разнообразны, чем конические сечения, и изучение их связано с бо́льшими трудностями. Мы рассмотрим бегло и без доказательств одну из самых интересных поверхностей этого типа: так называемый связный (или однополостный) гиперболоид.

Эта поверхность может быть получена следующим образом. Возьмем в пространстве три прямые l_1, l_2, l_3, находящиеся в общем положении. Последнее означает, что никакие две из них не параллельны и все три

не являются параллельными одной и той же плоскости. Может показаться удивительным, что существует бесконечное множество прямых в пространстве, из которых каждая пересекается со всеми тремя данными прямыми. Убедимся в этом. Пусть π — произвольная плоскость, содержащая прямую l_1; эта плоскость пересекает прямые l_2 и l_3 в двух точках, и прямая m, проведенная через эти две точки, очевидно, пересекается со всеми прямыми l_1, l_2 и l_3. Когда плоскость π вращается около прямой l_1, прямая m будет изменять свое положение, однако все время продолжая пересекаться с тремя данными прямыми. При движении m возникает поверхность, неограниченно уходящая в бесконечность, которая и называется однополостным гиперболоидом. Она содержит бесконечное множество прямых типа m. Любые три такие прямые, скажем m_1, m_2 и m_3, также будут находиться в общем положении, и те прямые в пространстве, которые будут пересекаться с тремя прямыми m_1, m_2 и m_3 одновременно, также будут лежать на рассматриваемой поверхности. Отсюда следует основное свойство гиперболоида: он составляется из двух различных семейств прямых линий; каждые три линии одного и того же семейства находятся в общем положении и каждая прямая одного семейства пересекается со всеми прямыми другого.

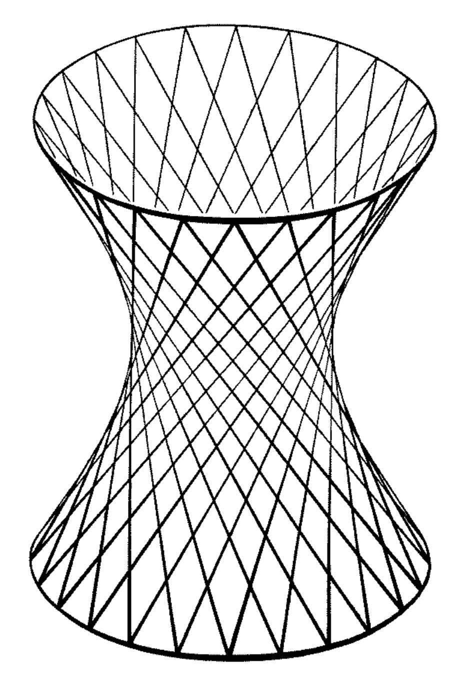


Рис. 109. Гиперболоид

Важное проективное свойство гиперболоида заключается в том, что двойное отношение тех четырех точек, в которых данная четверка прямых одного семейства пересекается с некоторой прямой второго семейства, не зависит от выбора этой последней. Это утверждение вытекает из метода построения гиперболоида с помощью вращающейся плоскости, и читатель может убедиться в его справедливости в качестве упражнения.

Отметим еще одно замечательное свойство гиперболоида: хотя он содержит два семейства прямых линий, но существование этих прямых не препятствует изгибанию поверхности — не делает ее жесткой. Если устроить модель гиперболоида из стержней, способных свободно вращаться около точек взаимных пересечений, то поверхность в целом

может быть непрерывно деформируема, пробегая бесконечное множество различных состояний.

§ 9. Аксиоматика и неевклидова геометрия

1. Аксиоматический метод. Аксиоматический метод в математике берет свое начало по меньшей мере от Евклида. Было бы совершенно ошибочно полагать, что античная математика развивалась или излагалась исключительно в строго постулативной форме, свойственной «Началам». Но впечатление, произведенное этим сочинением на последующие поколения, было столь велико, что в нем стали искать образцов для всякого строгого доказательства в математике. Иной раз даже философы (например, Спиноза в его «Ethica, more geometrico demonstrata») пытались излагать свои рассуждения в форме теорем, выводимых из определений и аксиом. В современной математике, после периода отхода от евклидовой традиции, продолжавшегося на протяжении XVII и XVIII вв., снова обнаружилось все усиливающееся проникновение аксиоматического метода в различные области. Одним из самых недавних продуктов подобного рода устремления мысли явилось возникновение новой дисциплины — математической логики.

В общих чертах аксиоматическая точка зрения может быть охарактеризована следующим образом. Доказать теорему в некоторой дедуктивной системе — значит установить, что эта теорема есть необходимое логическое следствие из тех или иных ранее доказанных предложений; последние в свою очередь должны быть доказаны и т. д. Процесс математического обоснования сводился бы, таким образом, к невыполнимой задаче «бесконечного спуска», если только в каком-нибудь месте нельзя было бы остановиться. Но в таком случае должно существовать некоторое число утверждений — *постулатов*, или *аксиом*, которые принимаются в качестве истинных и доказательство которых не требуется. Из них можно пытаться вывести все другие теоремы путем чисто логической аргументации. Если все факты некоторой научной области приведены в подобного рода логический порядок, а именно такой, что любой из них «выводится» из нескольких отобранных предложений (предпочтительно, чтобы таковые были немногочисленны, просты и легко усваивались), то тогда есть основание сказать, что область представима в «аксиоматической форме» или «допускает аксиоматизацию». Выбор предложений-аксиом в широкой степени произволен. Однако мало пользы, если наши постулаты недостаточно просты или если их слишком много. Далее, система постулатов должна быть *совместимой* (*непротиворечивой*) в том смысле, что никакие две теоремы, которые из них могут быть выведены, не должны содержать взаимных противоречий, и *полной* в том смысле, что всякая теорема, имеющая место в рассмат-

риваемой области, из них может быть выведена. Желательно также, чтобы система постулатов была *независимой*, т. е. чтобы ни один из них не был логическим следствием остальных. Вопрос о непротиворечивости и полноте системы аксиом был предметом больших дискуссий. Различные философские взгляды на источники человеческого знания обусловили различные, подчас несовместимые точки зрения на основания математики. Если математические понятия рассматриваются как субстанциальные объекты в сфере «чистой интуиции», независимые от определений и отдельных актов мыслительной деятельности человека, тогда, конечно, в математических результатах не может быть никаких противоречий, поскольку они представляют собой объективно истинные предложения, описывающие реальный мир. Если исходить из такой «кантианской» точки зрения, то никакой проблемы непротиворечивости вообще нет. Но, к сожалению, действительное содержание математики не удается уложить в столь простые философские рамки. Представители современного математического интуиционизма не полагаются на чистую интуицию в ее полном кантовском понимании. Они признают счетную бесконечность в качестве законного детища интуиции, но допускают использование лишь конструктивных свойств. Такие же фундаментальные понятия, как числовой континуум, следует, с их точки зрения, исключить из употребления, пожертвовав при этом важными разделами существующей математики (а то, что после этого остается, оказывается чрезвычайно сложным, причем без особой надежды на упрощение).

Совершенно другую позицию заняли «формалисты». Они не приписывают математическим понятиям никакой интуитивной реальности и не утверждают, что аксиомы выражают какие-то объективные истины, относящиеся к объектам чистой интуиции; они (формалисты) заботятся лишь о формальной логической правильности процесса рассуждений, базирующихся на постулатах. Позиция эта обладает безусловными преимуществами по сравнению с интуиционистской, так как она предоставляет математике полную свободу действий, нужную как для теории, так и для приложений. Но она вместе с тем вынуждает формалистов доказывать, что принятые ими аксиомы, выступающие теперь в качестве продукта свободного творчества человеческого интеллекта, не могут привести к противоречию. На протяжении последних двадцати лет[1] предпринимались многочисленные и напряженные попытки поиска такого рода доказательств непротиворечивости, особенно по отношению к аксиомам арифметики и алгебры и к понятию числового континуума. Результаты,

[1] Написано в 1941 г. О дальнейших работах в этой области, а также по поводу всей обширной проблематики оснований математики и характеристики различных направлений, см. [11], [15], [38]. — *Прим. ред.*

полученные в этом направлении, имеют исключительную важность, но задача в целом еще далеко не выполнена[2]. Более того, полученные в последние годы результаты свидетельствуют о том, что такого рода попытки и не могут привести к полному успеху — выяснилось, что для некоторых строго определенных и замкнутых систем понятий вообще нельзя доказать, что они непротиворечивы и в то же время полны. Особенно замечательно то обстоятельство, что все такого рода рассуждения, касающиеся проблем обоснования, проводятся полностью конструктивными и интуитивно убедительными методами.

Спор между интуиционистами и формалистами, особенно обострившийся в связи с парадоксами теории множеств (см. стр. 108–109), породил массу страстных выступлений убежденных сторонников обеих школ. Математический мир потрясали возгласы о «кризисе основ». Но эти сигналы тревоги не воспринимались — да и не следовало их воспринимать — слишком уж всерьез. При всем уважении к достижениям, завоеванным в борьбе за полную ясность основ, вывод, что эти расхождения во взглядах или же парадоксы, вызванные спокойным и привычным использованием понятий неограниченной общности, таят в себе серьезную угрозу для самого существования математики, представляется совершенно необоснованным.

Совершенно независимо от каких бы то ни было философских рассмотрений и интереса к проблемам оснований аксиоматический подход к предмету математики — самый естественный способ разобраться во всех хитросплетениях взаимосвязей между различными фактами и выяснить закономерности логического строения объединяющих их теорий. Не раз случалось, что такое сосредоточение внимания на формальной структуре, а не на интуитивном смысле понятий, облегчало отыскание обобщений и применений, которые легко было бы упустить при более интуитивном подходе к делу. Но выдающиеся открытия и подлинное понимание лишь в исключительных случаях оказывались результатом применения чисто аксиоматических методов. Подлинный источник развития математики — это творческая мысль, поддерживаемая интуицией. И если даже считать аксиоматизацию тем идеалом, к которому стремится математика, было бы непростительной ошибкой уверовать в то, что аксиоматика сама по себе является сутью математики. Творческая, конструктивная интуиция математика привносит в математику недедуктивные и иррациональные моменты, уподобляющие ее музыке или живописи.

Со времен Евклида геометрия неизменно была прототипом аксиоматизированной дисциплины. На протяжении столетий система евклидовых постулатов была предметом напряженного изучения. Но только

[2] См. предыдущее примечание. — *Прим. ред.*

сравнительно недавно стало совершенно ясно, что эти постулаты должны быть изменены и дополнены, для того чтобы из них могла быть выведена дедуктивно совокупность предложений элементарной геометрии. Например, в конце прошлого столетия Паш обнаружил, что при рассмотрении порядка расположения точек на прямой, т. е. соотношений, характеризуемых словом «между», требуется особый постулат. Паш выдвинул в качестве постулата следующее предложение: *если прямая пересекает сторону треугольника в точке, не являющейся вершиной, то она пересекается и еще с одной стороной треугольника.* (Невнимательное отношение к этой детали приводит к ряду явных парадоксов: абсурдные следствия — например, общеизвестное «доказательство» того, что все треугольники равнобедренные — как будто бы строго «выводятся» из евклидовых аксиом. Этот «вывод» основывается на неточном выполнении чертежа, причем некоторые прямые пересекаются вне треугольника или круга, тогда как на самом деле должны пересечься внутри.)

В своей знаменитой книге «Grundlagen der Geometrie» (первое издание ее появилось в 1899 г.) Гильберт дал вполне удовлетворительно построенную систему аксиом геометрии и вместе с тем произвел исчерпывающий анализ их взаимной независимости, их непротиворечивости и полноты.

Во всякую систему аксиом неизбежно входят некоторые неопределимые понятия, например, «точка» или «прямая» в геометрии. Их «значение» (или связь с объектами реального мира) *для математики* несущественно. Эти понятия должны быть принимаемы чисто абстрактно, и их математические свойства в пределах дедуктивной системы всецело вытекают из тех соотношений между ними, которые утверждаются в аксиомах. Так, в проективной геометрии естественно начать с основных понятий «точка» и «прямая» и отношения «инцидентности» и сформулировать две двойственные аксиомы: «каждые две различные точки инцидентны с одной и только одной прямой» и «каждые две различные прямые инцидентны с одной и только одной точкой». В аксиоматической системе проективной геометрии двойственность в формулировке аксиом обусловливает двойственность в самом построении. Всякой теореме, содержащей в своей формулировке и в доказательстве только двойственные элементы, непременно соответствует двойственная теорема. В самом деле, доказательство исходной теоремы заключается в последовательном применении некоторых аксиом, и применение в том же порядке двойственных аксиом составит доказательство двойственной теоремы.

Совокупность аксиом геометрии составляет *неявное определение* всех «неопределяемых» геометрических понятий: «точка», «прямая», «инцидентность» и т. д. Для применений геометрии важно, чтобы основные понятия и аксиомы геометрии находились в хорошем соответствии с

доступными физической проверке утверждениями, касающимися «реальных», осязаемых предметов. Физическая реальность, стоящая за понятием «точки», есть какой-то очень маленький объект, вроде небольшого пятнышка, получаемого на бумаге при прикосновении карандаша, и таким же образом «прямая» представляет собой абстракцию туго натянутой нити или светового луча. Свойства этих физических точек и прямых, как можно установить путем проверки, более или менее соответствуют формальным аксиомам геометрии. Легко себе представить, что более точно поставленные эксперименты могут вызвать необходимость в изменении аксиом, если мы хотим, чтобы они давали адекватное описание физических явлений. Напротив, если бы существовало заметное отклонение формальных аксиом от физических свойств предметов, то геометрия, построенная на этих аксиомах, представляла бы ограниченный интерес. Таким образом, даже с точки зрения формалиста, есть нечто, что оказывает большее влияние на направления математической мысли, нежели человеческий разум.

2. Гиперболическая неевклидова геометрия. В системе Евклида имеется одна аксиома, относительно которой — на основе сопоставления с эмпирическими данными, с привлечением туго натянутых нитей или световых лучей, — никак нельзя сказать, является ли она «истинной». Это знаменитый *постулат о параллельных*, утверждающий, что через данную точку, расположенную вне данной прямой, можно провести *одну и только одну* прямую, параллельную данной. Своеобразной особенностью этой аксиомы является то, что содержащееся в ней утверждение касается свойств прямой *на всем ее протяжении*, причем прямая предполагается неограниченно продолженной в обе стороны: сказать, что две прямые параллельны, — значит утверждать, что у них нельзя обнаружить общей точки, как бы далеко их ни продолжать. Вполне очевидно, что в пределах некоторой *ограниченной* части плоскости, как бы эта часть ни была обширна, напротив, можно провести через данную точку множество прямых, не пересекающихся с данной прямой. Так как максимально возможная длина линейки, нити, даже светового луча, изучаемого с помощью телескопа, непременно конечна, и так как внутри круга конечного радиуса существует много прямых, проходящих через данную точку и в пределах круга не встречающихся с данной прямой, то отсюда следует, что постулат Евклида никогда не может быть проверен экспериментально. Все прочие аксиомы Евклида имеют конечный характер, т. е. касаются конечных отрезков прямых или конечных частей рассматриваемых плоских фигур. Тот факт, что аксиома параллельности не допускает эмпирической проверки, выдвигает на первый план вопрос о том, является ли она *независимой* от прочих аксиом. Если бы она была неизбежным логическим следствием других

аксиом, то тогда нужно было бы просто вычеркнуть ее из списка аксиом и доказывать как теорему с помощью иных евклидовых аксиом. Много столетий математики пытались найти такое доказательство; этому способствовало широко распространенное среди всех, кто занимался геометрией, смутное сознание того, что аксиома параллельности по своему характеру существенно отличается от остальных, что ей недостает той убеждающей наглядности, которой, казалось бы, должно было обладать всякое геометрическое предложение, возводимое в ранг аксиомы.

Одна из первых попыток в указанном направлении была сделана в IV столетии н. э. комментатором Евклида Проклом, который, чтобы избежать необходимости вводить специальный постулат о параллельных прямых, ввел *определение*, согласно которому прямая, параллельная данной прямой, есть геометрическое место точек, расположенных от нее на одном и том же заданном расстоянии. При этом Прокл упустил из виду, что таким образом трудность не устраняется, а только перемещается, так как при его ходе мыслей остается недоказанным, что названное геометрическое место действительно есть прямая линия. Так как последнего Прокл доказать не мог, то именно это предложение ему пришлось бы принять в качестве аксиомы параллельности, и ничто не было бы выиграно, так как мы можем легко установить, что обе упомянутые аксиомы эквивалентны между собой. Саккери (1667–1733), а затем Ламберт (1728–1777) делали попытки доказать аксиому параллельности косвенным путем, допуская противоположное утверждение и выводя из него абсурдные следствия. Но выведенные ими следствия оказались далеко не абсурдными: это были теоремы неевклидовой геометрии, получившей позднее дальнейшее развитие. Если бы названные лица рассматривали свои результаты не как нелепости, а как утверждения, свободные от внутренних противоречий, то не кому иному, как им, принадлежала бы заслуга открытия неевклидовой геометрии.

Но в те времена любую геометрическую систему, не находящуюся в абсолютном согласии с евклидовой, непременно стали бы рассматривать как очевидную нелепость. Кант, наиболее влиятельный философ той эпохи, выразил свое отношение к вопросу, утверждая, что аксиомы Евклида — не что иное, как неизбежные формы человеческого мышления, чем, по его мнению, и объясняется их объективная значимость по отношению к «реальному» пространству. Эта вера в аксиомы Евклида, как в незыблемые истины, существующие в сфере чистой интуиции, была одним из главных догматов кантовой философии. Однако с течением времени ни привычные навыки мышления, ни влияние философских авторитетов не смогли подавить растущего убеждения, что неизменные неудачи в поисках доказательства аксиомы параллельности имели своей причиной не столько недостаток изобретательности со стороны геометров, сколько тот основной факт, что этот постулат на самом деле

независим от других. (Подобным же образом неудачи в решении при помощи радикалов общего уравнения пятой степени мало-помалу привели к подозрению, позднее оправдавшемуся, что такое решение невозможно.) Венгерский математик Бойяи (1802–1860) и русский математик Лобачевский (1793–1856) положили конец сомнениям, построив во всех деталях геометрическую систему, в которой аксиома параллельности была отвергнута. Когда молодой гениальный энтузиаст Бойяи послал свою работу «королю математики» Гауссу, от которого с нетерпением ждал поддержки, то получил в ответ уведомление, что самим Гауссом открытие было сделано раньше, но он воздержался в свое время от публикации результатов, опасаясь слишком шумных обсуждений.

Посмотрим, что же означает независимость аксиомы параллельности. Эту независимость следует понимать в том смысле, что возможно свободное от внутренних противоречий построение «геометрических» предложений о точках, прямых и т. д., исходя из системы аксиом, в которой аксиома параллельности заменена противоположной. Такое построение называется неевклидовой геометрией. Нужно было интеллектуальное бесстрашие Гаусса, Бойяи и Лобачевского, чтобы осознать, что геометрия, основанная не на евклидовой системе аксиом, может быть абсолютно непротиворечивой.

Чтобы убедиться в непротиворечивости новой геометрии, нет надобности развивать во всех подробностях многочисленные теоремы неевклидовой геометрии, как это делали Бойяи и Лобачевский. Мы умеем теперь строить простые «модели» такой геометрии, удовлетворяющие всем аксиомам Евклида, кроме аксиомы параллельности. Простейшая модель была указана Феликсом Клейном, работы которого в этой области стимулировались идеями английского геометра Кэли (1821–1895). В такой модели через данную точку, лежащую вне данной прямой, можно провести бесчисленное множество «прямых», «параллельных» данной прямой. Подобного рода геометрия называется геометрией Бойяи—Лобачевского, или «гиперболической» геометрией. (Основание для последнего наименования будет приведено на стр. 246.)

При построении модели Клейна сначала рассматриваются объекты обыкновенной евклидовой геометрии; и затем некоторые из объектов и отношений между ними *переименовываются* таким образом, что для их описания оказывается пригодной уже неевклидова геометрия. Эта последняя, тем самым, не в меньшей мере непротиворечива, чем первоначальная евклидова геометрия, так как излагается (если посмотреть с другой точки зрения и описывать другими словами) как совокупность фактов обыкновенной евклидовой геометрии. С этой моделью можно легко освоиться, привлекая кое-какие понятия из проективной геометрии.

При проективном преобразовании одной плоскости на другую или на

саму себя (можно после отображения совместить обе плоскости) окружность, вообще говоря, переходит в некоторое коническое сечение. Но можно легко показать (доказательства мы не приводим), что существует бесчисленное множество таких проективных преобразований плоскости на саму себя, при которых данный круг, вместе со всеми заключенными внутри точками, переходит сам в себя. При таких преобразованиях внутренние точки, как и точки контура, меняют, вообще говоря, свои места, но внутренние точки остаются внутренними, а точки контура остаются на контуре. (Центр круга, как легко убедиться, можно перевести в любую наперед заданную внутреннюю точку.) Рассмотрим совокупность всех таких преобразований. Конечно, они не будут оставлять очертания фигур неизменными и потому не являются движениями в обычном смысле. Но мы теперь сделаем решающий шаг и *назовем* их «неевклидовыми движениями» в той геометрии, которую строим. Посредством этих «движений» можно дальше определить и «равенство»: две фигуры *называются* равными, если существует «неевклидово движение», переводящее одну фигуру в другую.

Перейдем теперь к описанию упомянутой выше клейновой модели гиперболической геометрии. «Плоскость» состоит только из внутренних точек круга, внешние точки просто отбрасываются. Каждая внутренняя точка *называется* неевклидовой «точкой», каждая хорда круга *называется* неевклидовой «прямой»; «движения» и «равенства» уже определены выше; проведение «прямой» через две «точки» и нахождение «точки» пересечения двух «прямых» совершаются, как в евклидовой геометрии. Легко убедиться, что новая конструкция удовлетворяет всем постулатам евклидовой геометрии, с единственным исключением — постулатом о параллельных прямых. Что этот постулат здесь не выполняется, ясно видно из того, что через «точку», не лежащую на «прямой», можно провести бесчисленное множество «прямых», не имеющих общей «точки» с данной прямой. Данная «прямая» есть евклидова хорда, тогда как в качестве второй «прямой» может быть взята любая из хорд, проходящих через данную «точку» и не пересекающих первой «прямой» внутри круга. Описанная простая модель совершенно достаточна для того, чтобы покончить с основным вопросом, породившим неевклидову геометрию: она показывает, что аксиома параллельности не выводится из остальных аксиом евклидовой геометрии. Действительно, если бы она выводилась из них, то тогда была бы верной теоремой и по

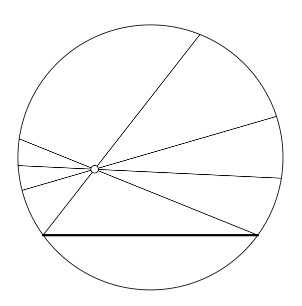


Рис. 110. Модель неевклидовой плоскости Клейна

отношению к модели Клейна, а мы видим, что это не так.

Строго говоря, предыдущая аргументация построена на допущении, что модель Клейна непротиворечива, т. е. что нельзя доказать вместе с некоторым утверждением также и противоположного утверждения. Но, во всяком случае, геометрия модели Клейна непротиворечива в такой же степени, как и обыкновенная евклидова геометрия, так как теоремы о «точках» и «прямых» и т. д. модели Клейна представляют собой только своеобразно сформулированные теоремы евклидовой геометрии. Удовлетворительного доказательства непротиворечивости аксиом евклидовой геометрии дано не было, если не считать сведения к аналитической геометрии и в конечном счете к числовому континууму; а непротиворечивость концепции континуума — также вопрос открытый[1].

* Мы привлечем внимание читателя еще к одной детали (впрочем, стоящей за пределами непосредственно поставленных нами задач) — именно, к определению неевклидова «расстояния» в модели Клейна. Это «расстояние» должно быть инвариантно относительно неевклидовых «движений», так как обыкновенное движение не изменяет обыкновенного расстояния. Мы знаем, что двойное отношение есть инвариант проективного преобразования. Естественно возникает мысль о том, чтобы при определении «расстояния» между двумя различными точками P и Q внутри нашего круга воспользоваться двойным отношением $(OSQP)$, где O и S — точки, в которых продолженный в обе стороны отрезок PQ встречается с окружностью. Это двойное отношение, в самом деле, есть положительное число; но взять это отношение непосредственно в качестве «расстояния» PQ не представляется удобным. Действительно, в предположении, что три точки P, Q, R лежат на одной прямой, мы должны были бы иметь равенство $\overline{PQ}+\overline{QR}=\overline{PR}$, но, вообще говоря,

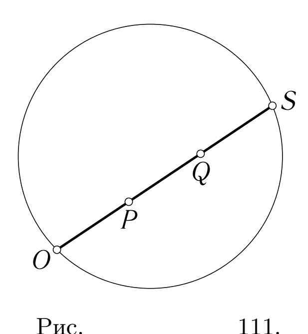


Рис. 111. Неевклидово расстояние

$$(OSQP)+(OSRQ)\neq(OSPR).$$

Напротив, справедливо несколько иное равенство

$$(OSQP)\cdot(OSRQ)=(OSPR); \tag{1}$$

в самом деле,

$$(OSQP)\cdot(OSRQ)=\left\{\frac{QO}{QS}:\frac{PO}{PS}\right\}\cdot\left\{\frac{RO}{RS}:\frac{QO}{QS}\right\}=\frac{RO}{RS}:\frac{PO}{PS}=(OSRP).$$

[1] Подробнее об исследованиях в этой области см. упомянутую на стр. 109 книгу А. Френкеля и И. Бар-Хиллела, содержащую также обширную библиографию. — *Прим. ред.*

Свойство (1) позволяет определить «расстояние» PQ как *логарифм двойного отношения* (а не как само двойное отношение), с таким расчетом, чтобы обеспечить аддитивность расстояния: $\overline{PQ}$ = неевклидово «расстояние» $PQ = \log(OSQP)$. Это «расстояние» есть положительное число, так как $(OSQP) > 1$ при $P \neq Q$.

Из основного свойства логарифма (см. стр. 467) следует, в силу (1), что $\overline{PQ} + \overline{QR} = \overline{PR}$. По какому основанию брать логарифмы — несущественно, так как при изменении основания меняется лишь единица измерения. Между прочим, если одна из точек, скажем Q, приближается к окружности, то неевклидово расстояние PQ неограниченно возрастает. Это означает, что «прямая» нашей неевклидовой модели имеет бесконечную неевклидову «длину», хотя в евклидовом смысле представляет собой конечный отрезок.

3. Геометрия и реальность. Модель Клейна показывает, что гиперболическая геометрия как формально-дедуктивное построение непротиворечива в такой же степени, как и классическая евклидова геометрия. Возникает вопрос: которой же из двух геометрий следует отдать предпочтение, когда речь идет об описании геометрических отношений, существующих в физическом мире? Как мы уже отметили, эксперимент никоим образом не может решить, проходит ли через данную точку только одна прямая, параллельная данной прямой, или бесчисленное множество. Однако в евклидовой геометрии сумма углов треугольника равна 180°, тогда как в гиперболической геометрии, как можно показать, она меньше 180°. Гаусс предпринял опытное исследование вопроса о том, как обстоит дело с суммой углов треугольника с физической точки зрения: он очень тщательно измерил углы в треугольнике, образованном тремя достаточно удаленными друг от друга горными пиками, и в пределах возможных ошибок измерений сумма углов оказалась равной 180°. Если бы результат был заметно меньше 180°, то отсюда следовало бы, что гиперболическая геометрия лучше подходит для описания внешнего мира. Но эксперимент не решил ничего, так как для небольших треугольников со сторонами длиной всего в несколько миль отклонение от 180°, которое предвидит гиперболическая геометрия, могло быть столь ничтожным, что гауссовы инструменты его не обнаружили. Таким образом, не дав решающих результатов, эксперимент все же показал, что евклидова и гиперболическая геометрии, различающиеся только в очень *обширных* частях пространства, для сравнительно малых фигур оказываются практически одинаково пригодными для употребления. Поэтому если рассматриваются только *локальные* свойства пространства, то выбор между двумя геометриями остается делать лишь по принципу простоты. Но так как работать с евклидовой геометрией гораздо легче, чем с ги-

перболической, то мы и пользуемся именно ею, покуда рассматриваются небольшие (порядка нескольких миллионов миль!) расстояния. Однако нет оснований ожидать, что она наверное оказалась бы подходящей при описании физического мира в целом, во всех его обширных пространствах. Положение вещей в геометрии совершенно такое же, какое существует и в физике, где системы Ньютона и Эйнштейна дают неразличимые результаты при малых расстояниях и скоростях, но обнаруживают расхождение, когда рассматриваются большие величины.

Научно-революционное значение открытия неевклидовой геометрии заключается в том, что оно разрушило представление об аксиомах Евклида как о непоколебимой математической схеме, к которой приходится приспособлять наши экспериментальные знания о физической реальности.

4. Модель Пуанкаре. Математик волен видеть «геометрию» во всякой непротиворечивой системе аксиом, говорящих о «точках», «прямых» и т. д.; но его исследования только в том случае будут полезны для физика, если система аксиом находится в соответствии с поведением физических объектов в реальном мире. Мы хотели бы теперь, с этой точки зрения, разобраться в смысле утверждения: «Свет распространяется по прямой линии». Если в этом утверждении содержится *физическое определение* «прямой линии», то систему геометрических аксиом следует выбирать таким образом, чтобы получилось соответствие с поведением световых лучей. Вообразим, следуя Пуанкаре, что мир состоит из внутренности круга C и что во всякой точке скорость света пропорциональна расстоянию точки от окружности. Можно тогда доказать, что свет будет распространяться по круговым дугам, образующим прямые углы с окружностью C. В таком мире геометрические свойства «прямых линий» (определенных как световые лучи) будут отличаться от свойств евклидовых прямых. В частности, не будет евклидовой аксиомы параллельности, так как через данную точку пройдет бесчисленное множество «прямых линий», не пересекающихся с данной «прямой линией». Можно обнаружить, что «точки» и «прямые линии» в описываемом мире будут обладать в точности теми же свойствами, какими обладают

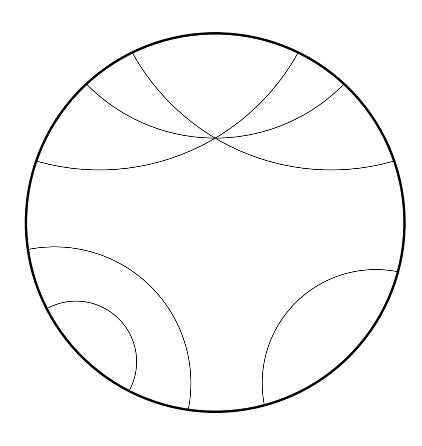


Рис. 112. Модель неевклидовой плоскости Пуанкаре

«точки» и «прямые» в модели Клейна. Другими словами, мы получили новую модель гиперболической геометрии. Но евклидову геометрию также можно применять в этом мире: тогда выйдет, что световые лучи, которые уже не будут евклидовыми «прямыми линиями», распространяются по кругам, перпендикулярным к окружности C. Таким образом, одна и та же физическая ситуация может быть описана различными геометрическими системами, если предположить, что физические объекты (в нашем случае — световые лучи) связаны с различными понятиями в этих системах:

Световой луч → «прямая линия» — гиперболическая геометрия
Световой луч → «окружность» — евклидова геометрия

Так как в евклидовой геометрии понятие прямой линии сопоставляется с поведением светового луча в однородной среде, то говоря, что геометрия в описании мира внутри C гиперболическая, мы утверждали бы только то, что физические свойства световых лучей в этом мире те же самые, что и свойства «прямых» гиперболической геометрии.

5. Эллиптическая, или риманова, геометрия. В евклидовой геометрии, как и в гиперболической геометрии Бойяи—Лобачевского, молчаливо допускается, что всякая прямая бесконечна (бесконечность прямой существенно связана с отношением «быть между» и аксиомами порядка). Но, после того как гиперболическая геометрия открыла путь к свободному построению геометрий, естественно возник вопрос о том, нельзя ли осуществить построение таких неевклидовых геометрий, в которых прямые линии конечны и замкнуты. Разумеется, в таких геометриях теряют силу не только постулат о параллельных, но и аксиомы порядка. Современные исследования выяснили значение этих геометрий для новейших физических теорий. Впервые такие геометрии были подвергнуты рассмотрению в речи, произнесенной в 1851 г. Риманом при вступлении его в должность приват-доцента Гёттингенского университета. Геометрии с замкнутыми конечными прямыми могут быть построены без каких бы то ни было противоречий. Вообразим двумерный мир, состоящий из поверхности S сферы, причем под «прямыми» условимся понимать большие круги сферы. Это был бы самый естественный способ описывать «мир» мореплавателя: дуги больших кругов являются кратчайшими кривыми, связывающими две

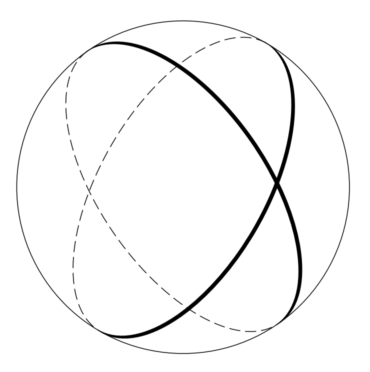


Рис. 113. «Прямые линии» в геометрии Римана

точки на сфере, а это как раз и есть характеристическое свойство прямых на плоскости. В рассматриваемом двумерном мире *всякие* две «прямые» пересекаются, так что из внешней точки нельзя провести *ни одной* «прямой», не пересекающейся с данной (т. е. ей параллельной). Геометрия «прямых» в этом мире называется *эллиптической геометрией*. Расстояние между двумя точками в такой геометрии измеряется просто как длина кратчайшей дуги большого круга, проходящего через данные точки. Углы измеряются так же, как и в евклидовой геометрии. Самым характерным свойством эллиптической геометрии мы считаем несуществование параллельных.

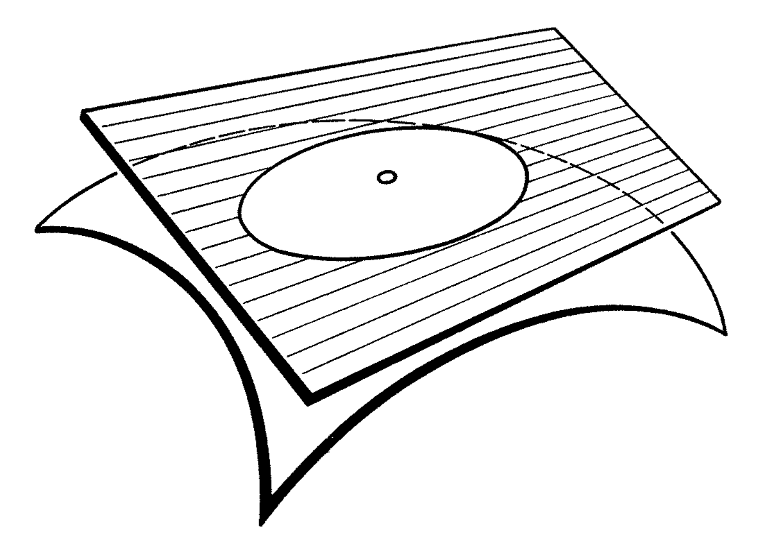


Рис. 114. Эллиптическая точка

Следуя Риману, мы можем обобщить эту геометрию следующим образом. Рассмотрим «мир», состоящий из некоторой кривой поверхности в пространстве (не обязательно сферы) и определим «прямую линию», проходящую через две точки, как *кратчайшую кривую* («геодезическую»), соединяющую эти точки. Точки поверхности можно разбить на два класса: 1°. Точки, в окрестности которых поверхность подобна сфере в том отношении, что она вся лежит по одну сторону от касательной плоскости в этой точке. 2°. Точки, в окрестности которых поверхность седлообразна — лежит по обе стороны касательной плоскости. Точки первого класса называются эллиптическими точками поверхности — по той причине, что при небольшом параллельном перемещении касательной плоскости она пересечет поверхность по кривой, имеющей вид эллипса; точки же второго класса носят название гиперболических, так как при аналогичном перемещении касательной плоскости полу-

чается пересечение с поверхностью, напоминающее гиперболу. Геометрия геодезических «прямых» в окрестности точки поверхности является эллиптической или гиперболической, смотря по тому, будет ли сама точка эллиптической или гиперболической. В этой модели неевклидовой геометрии углы измеряются, как в обыкновенной евклидовой геометрии.

Изложенная идея была развита Риманом дальше: он рассмотрел геометрии пространства, аналогичные только что разобранным геометриям поверхности. По Риману, «кривизна» пространства, меняясь от точки к точке, определяет характер геометрии в окрестности точки. «Прямые линии» у Римана — геодезические кривые. В эйнштейновой общей теории относительности геометрия пространства есть риманова геометрия; свет распространяется по геодезическим линиям, а кривизна пространства в каждой точке определяется в зависимости от свойств материи в окрестности точки.

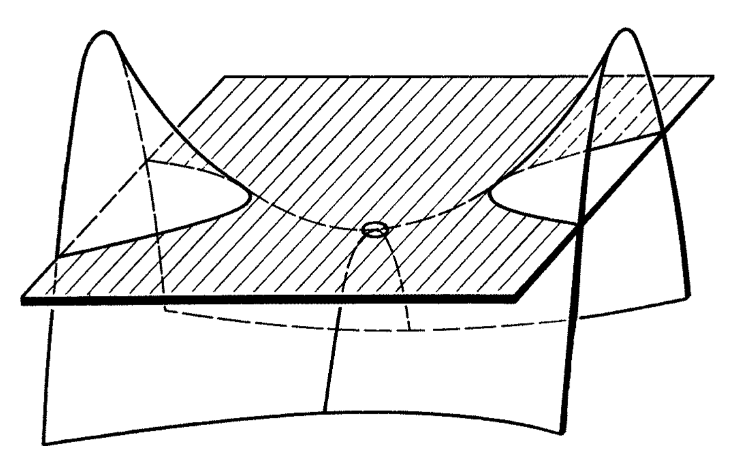


Рис. 115. Гиперболическая точка

Возникнув из чисто аксиоматических изысканий, неевклидова геометрия в наши дни стала чрезвычайно полезным аппаратом, допускающим различные применения при изучении физической реальности. В теории относительности, в оптике, в общей теории колебаний неевклидово описание явлений оказывается в ряде случаев гораздо более адекватным физической реальности, чем евклидово.

ПРИЛОЖЕНИЕ

Геометрия в пространствах более чем трех измерений

1. Введение. То «реальное» пространство, которое служит средой нашего физического опыта, имеет три измерения, плоскость имеет два измерения, прямая — одно. Наша, в обычном смысле понимаемая, пространственная интуиция решительно ограничена тремя измерениями — и дальше не простирается. Тем не менее во многих случаях вполне уместно говорить о «пространствах», имеющих четыре или более измерений. В каком же смысле допустимо говорить об n-мерном пространстве, где $n > 3$, и для чего могут быть полезны такие пространства? Ответ можно дать, став или на аналитическую, или на геометрическую точку зрения. Терминологию n-мерного пространства дозволительно рассматривать только как образный язык, служащий для выражения математических идей, находящихся за пределами обычной геометрической интуиции.

2. Аналитический подход. Мы уже обращали внимание читателя на изменение роли аналитической геометрии, происшедшее на протяжении ее развития. Точки, прямые, кривые линии и т. д. первоначально рассматривались как чисто геометрические объекты, и задачей аналитической геометрии было всего-навсего, сопоставляя им координаты или уравнения, интерпретировать и развивать дальше геометрическую теорию алгебраическими или аналитическими методами. Но с течением времени постепенно начала утверждаться противоположная точка зрения. Число x, или пара чисел x, y, или тройка чисел x, y, z стали рассматриваться как исходные, основные объекты, и эти аналитические объекты далее конкретизировались, или, еще лучше сказать, «визуализировались» в виде точек на прямой, на плоскости, в пространстве. И тогда геометрический язык стал служить для того, чтобы констатировать наличие тех или иных соотношений между числами. При этом мы лишаем геометрические объекты их самостоятельного и независимого значения и говорим, что пара чисел x, y *есть* точка на плоскости, совокупность всех пар x, y, удовлетворяющих линейному уравнению $L(x, y) = ax + by + c = 0$ (где a, b, c — данные постоянные числа), *есть* прямая линия и т. д. Такие же определения устанавливаются и для трехмерного пространства.

Даже в том случае, когда мы занимаемся собственно алгебраической проблемой, язык геометрии нередко представляется вполне удобным для краткого и совершенно точного описания фактов, и геометрическая интуиция начинает работать, подсказывая правильные алгебраи-

ческие процедуры. Например, решая систему трех линейных уравнений с тремя неизвестными x, y, z

$$\left.\begin{aligned} L(x,\,y,\,z) = ax + by + cz + d = 0 \\ L'(x,\,y,\,z) = a'x + b'y + c'z + d' = 0 \\ L''(x,\,y,\,z) = a''x + b''y + c''z + d'' = 0 \end{aligned}\right\},$$

мы истолковываем стоящую перед нами задачу геометрически и говорим, что в трехмерном пространстве R_3 требуется найти точку пересечения трех плоскостей, заданных уравнениями $L = 0$, $L' = 0$, $L'' = 0$. Другой пример: рассматривая все такие числовые пары x, y, что $x > 0$, мы скажем, что имеем дело с полуплоскостью, расположенной вправо от оси y. В более общем случае совокупность числовых пар x, y, для которых выполняется неравенство

$$L(x,\,y) = ax + by + c > 0,$$

интерпретируется как полуплоскость, лежащая по одну сторону прямой $L = 0$, а совокупность таких числовых троек x, y, z, что

$$L(x,\,y,\,z) = ax + by + cz + d > 0,$$

— как «полупространство», определяемое плоскостью $L = 0$.

После этих разъяснений нам совсем легко перейти к «четырехмерному» или даже к «n-мерному» пространству. Рассмотрим четверку чисел x, y, z, t. Скажем, что такая четверка представляет собой точку, или, еще проще, *есть* точка в четырехмерном пространстве R_4. Вообще, по определению, точка n-мерного пространства R_n есть не что иное, как система из n действительных чисел $x_1, x_2, \ldots, x_n$, записанных в определенном порядке. Не так важно, что мы не «видим» этой точки. Геометрический язык не перестает быть вполне понятным в случае, если идет речь об алгебраических свойствах n переменных. Дело в том, что многие алгебраические свойства линейных уравнений и т. п. совершенно не зависят от числа входящих переменных, или, как принято говорить, от размерности пространства этих переменных. Мы назовем, таким образом, «гиперплоскостью» совокупность всех таких точек $x_1, x_2, \ldots, x_n$ в n-мерном пространстве R_n, которые удовлетворяют линейному уравнению

$$L(x_1,\,x_2,\,\ldots,\,x_n) = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b = 0.$$

Точно так же основная алгебраическая задача решения системы n линейных уравнений с n неизвестными

$$\left.\begin{aligned} L_1(x_1,\,x_2,\,\ldots,\,x_n) = 0 \\ L_2(x_1,\,x_2,\,\ldots,\,x_n) = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ L_n(x_1,\,x_2,\,\ldots,\,x_n) = 0 \end{aligned}\right\}$$

истолковывается на геометрическом языке как нахождение точки пересечения n гиперплоскостей $L_1 = 0$, $L_2 = 0$, …, $L_n = 0$.

Преимущество такого геометрического способа описания математических факты заключается в том, что он подчеркивает некоторые обстоятельства алгебраического характера, которые не зависят от числа измерений n и вместе с тем в случае $n \leqslant 3$ могут быть наглядно интерпретированы. Во многих приложениях употребление геометрической терминологии имеет также преимущество краткости, и вместе с тем облегчает аналитические рассуждения, а иногда руководит ими и направляет их в должную сторону. Теория относительности снова может быть приведена здесь в качестве примера области, в которой существенный успех был достигнут по той причине, что три пространственные координаты x, y, z и временная координата t «события» были объединены в одно «пространственно-временнóе» четырехмерное многообразие x, y, z, t. Подчиняя, таким образом, «пространство-время» этой аналитической схеме и наделяя его, кроме того, свойствами неевклидовой геометрии, удалось описать многие весьма сложные ситуации с замечательной простотой. Столь же полезными оказались n-мерные пространства в механике, в статистической физике, не говоря уже о самой математике.

Приведем еще кое-какие чисто математические примеры. Совокупность всех кругов на плоскости образует трехмерное многообразие, так как круг с центром x, y и радиусом t может быть изображен точкой с координатами x, y, t. Так как радиус круга есть положительное число, то совокупность рассматриваемых точек заполняет полупространство. Таким же образом совокупность всех сфер в обыкновенном трехмерном пространстве образует четырехмерное многообразие, так как каждая сфера с центром x, y, z и радиусом t может быть представлена точкой с координатами x, y, z, t. Куб в трехмерном пространстве с центром в начале координат, ребрами длины 2 и гранями, параллельными координатным плоскостям, состоит из совокупности всех точек x_1, x_2, x_3, для которых $|x_1| \leqslant 1$, $|x_2| \leqslant 1$, $|x_3| \leqslant 1$. Так же точно «куб» в n-мерном пространстве R_n с центром в начале координат, «ребрами» длины 2 и «гранями», параллельными координатным плоскостям, определяется как совокупность точек x_1, x_2, …, x_n, для которых одновременно справедливы неравенства $|x_1| \leqslant 1$, $|x_2| \leqslant 1$, …, $|x_n| \leqslant 1$. «Поверхность» такого куба состоит из всех точек, для которых хотя бы в одном из этих соотношений имеет место знак равенства. Поверхностные элементы размерности $n-2$ состоят из точек, для которых знак равенства стоит по меньшей мере два раза; и т. д.

Упражнение. Дайте описание поверхности такого куба в трехмерном, четырехмерном, n-мерном пространствах.

***3. Геометрический, или комбинаторный, подход.** Хотя аналитический подход к n-мерной геометрии чрезвычайно прост и удобен для многих приложений, все же следует упомянуть и о другом методе, носящем чисто геометрический характер. Он основан на редукции от n-мерных данных к $(n-1)$-мерным и тем открывает возможность определять многомерные геометрии посредством математической индукции.

Начнем с того, что рассмотрим контур треугольника ABC в двух измерениях. Разрезая его в точке C и затем поворачивая стороны AC и BC соответственно около A и B, мы выпрямим контур в прямолинейный отрезок (рис. 116), на котором точка C будет фигурировать дважды. Полученная одномерная фигура дает исчерпывающее представление контура двумерного треугольника. Сгибая фигуру в точках A и B и добившись совпадения двух точек C, мы имеем возможность восстановить треугольник. Но важно то, что сгибать вовсе и не нужно. Достаточно условиться, что мы «идентифицируем» (т. е. не будем различать) обе точки C, несмотря на то что эти две точки и не совпадают в обычном смысле. Можно сделать еще следующий шаг: разрезая фигуру также и в точках A и B, мы получим три отрезка CA, AB, BC, которые при желании можно опять сложить таким образом, чтобы был восстановлен «настоящий» треугольник ABC, причем пары иден-

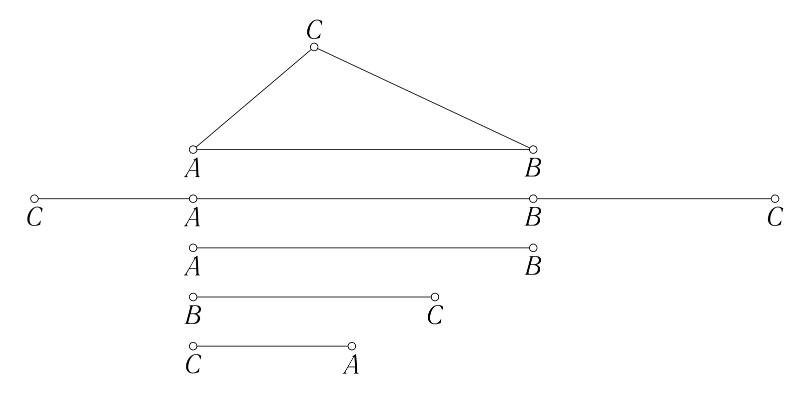


Рис. 116. Определение треугольника по сторонам с сопоставленными друг другу концами

тифицируемых точек совпадут между собой. Идея идентифицировать различные точки в данной совокупности отрезков, чтобы из них построить многоугольный контур (в нашем случае — треугольник), практически иногда оказывается очень полезной. Если нужно отправить в дальнее путешествие какое-нибудь соединение из металлических балок, например, мостовую ферму, то удобнее всего упаковать сложенные вместе, предварительно разъединенные балки, обозначив одними и теми

же знаками те концы различных балок, которые должны быть соединены вместе. Такое собрание балок с размеченными концами совершенно эквивалентно пространственной конструкции. Предыдущее замечание приводит к мысли о том, как можно «разнять» двумерный многогранник в трехмерном пространстве, заменяя его фигурами низших измерений. Возьмем, например, поверхность куба (рис. 117). Ее сейчас же можно свести к системе из шести квадратов, стороны которых надлежащим образом идентифицированы; следующий шаг будет состоять в том, чтобы заменить эту систему квадратов системой из 12 прямолинейных отрезков с надлежащим образом идентифицированными концами.

Вообще, любой многогранник в трехмерном пространстве R_3 приводится таким образом или к системе плоских многоугольников, или к системе прямолинейных отрезков.

Упражнение. Выполните указанную редукцию для всех правильных многогранников (см. стр. 257).

Теперь уже ясно, что мы можем обратить ход наших рассуждений, *определяя* многоугольник на плоскости с помощью системы прямолинейных отрезков и многогранник в пространстве R_3 — с помощью системы многоугольников в R_2 или же, при условии дальнейшей редукции, с помощью опять-таки прямолинейных отрезков. Но тогда совершенно естественно определить «многогранник» в четырехмерном пространстве R_4 с помощью системы многогранников в R_3 при надлежащей идентификации двумерных граней; «многогранник» в R_5 — с помощью «многогранников» в R_4 и т. д. В конечном счете всякий «многогранник» в R_n сводится к системе отрезков.

Останавливаться на этом вопросе подробнее мы лишены возможности. Добавим лишь несколько замечаний, не приводя доказательств. «Куб» в R_4 ограничен 8 трехмерными кубами, из которых каждый имеет со своими «соседями» по идентифицированной двумерной грани. У такого куба 16 вершин, в каждой вершине сходятся по четыре ребра; всего ребер имеется 32. В R_4 существует шесть правильных многогранников. Кроме «куба», имеется один многогранник, ограниченный 5 правильными тетраэдрами, один, ограниченный 16 тетраэдрами, один, ограниченный 24 октаэдрами, один, ограниченный 120 додекаэдрами, и еще один, ограниченный 600 тетраэдрами. Доказано, что в R_n, при $n > 4$, существует только 3 правильных многогранника: один с $n+1$ вершинами, ограниченный $n+1$ многогранниками из R_{n-1}, имеющими по n $(n-2)$-мерных граней; один с 2^n вершинами, ограниченный $2n$ многогранниками из R_{n-1}, имеющими по $2n-2$ $(n-2)$-мерных граней; и еще один с $2n$ вершинами, ограниченный 2^n многогранниками из R_{n-1}, имеющими по n $(n-2)$-мерных граней.

Упражнение. Сравните определение «куба» из R_4, данное в пункте 2, с

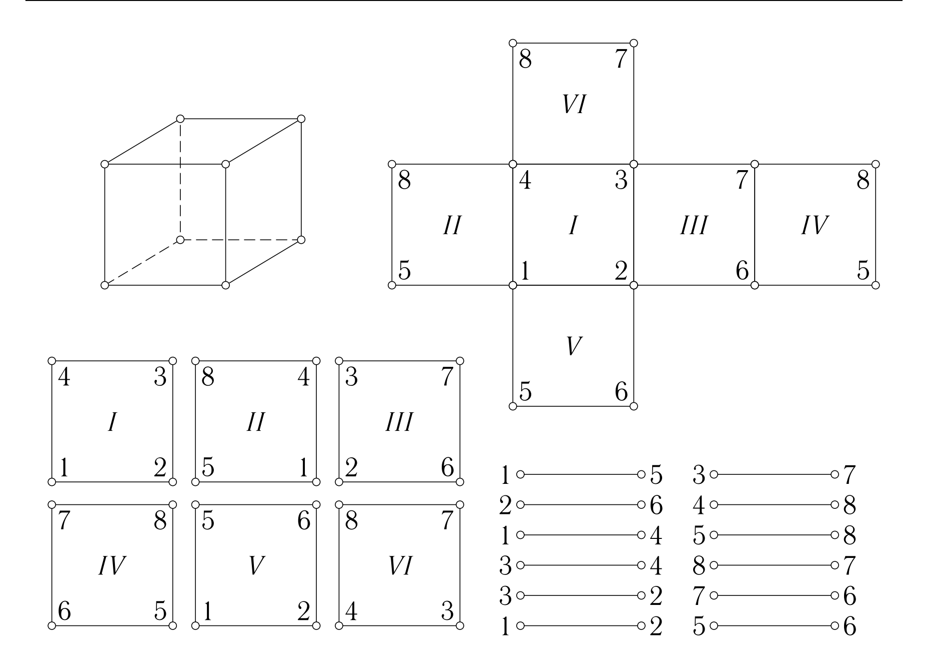


Рис. 117. Определение куба по сопоставленным друг другу вершинам и ребрам

определением, данным в настоящем пункте, и установите, что прежнее «аналитическое» определение куба равносильно настоящему «комбинаторному».

Со структурной, или «комбинаторной», точки зрения простейшими геометрическими фигурами размерности 0, 1, 2, 3 являются соответственно точка, отрезок, треугольник, тетраэдр. Ради единообразия символики обозначим фигуры этого типа соответственно T_0, T_1, T_2, T_3. (Индексы указывают на размерность.) Структура каждой из этих фигур характеризуется тем, что каждая фигура типа T_n имеет $n+1$ вершин и каждое подмножество из $i+1$ вершин фигуры типа T_n $(i=0,\ 1,\ \ldots,\ n)$ определяет некоторую фигуру типа T_i. Например, трехмерный тетраэдр T_3 имеет 4 вершины, 6 ребер и 4 грани.

Ясно, как будет дальше. Мы определим четырехмерный «тетраэдр» T_4 как множество, состоящее из 5 вершин, причем каждое подмножество из 4 вершин порождает фигуру типа T_3, каждое подмножество из 3 вершин — фигуру типа T_2 и т. д. Фигура типа T_4 схематически показана на рис. 118: мы видим, что у нее 5 вершин, 10 ребер, 10 треугольных граней и 5 тетраэдров.

Обобщение на n измерений не представляет труда. Из теории соединений известно, что существует ровно $C_i^r = \dfrac{r!}{i!\,(r-i)!}$ таких различных подмножеств по i объектов, которые могут быть составлены из множе-

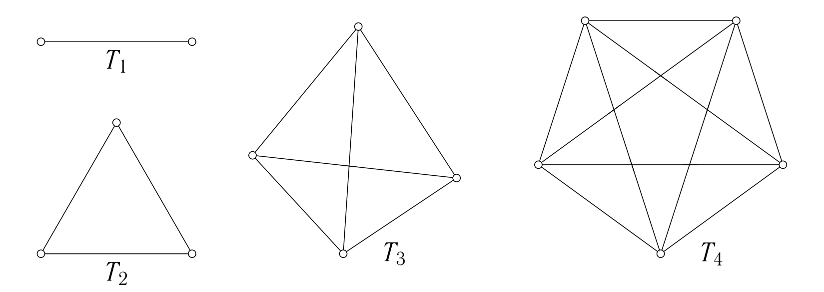


Рис. 118. Простейшие элементы в 1, 2, 3, 4 измерениях

ства r объектов. Поэтому n-мерный «тетраэдр» содержит

$C_1^{n+1} = n+1$ вершин (фигур типа T_0),

$C_2^{n+1} = \dfrac{(n+1)!}{2!\,(n-1)!}$ ребер (фигур типа T_1),

$C_3^{n+1} = \dfrac{(n+1)!}{3!\,(n-2)!}$ треугольников (фигур типа T_2),

$C_4^{n+1} = \dfrac{(n+1)!}{4!\,(n-3)!}$ фигур типа T_3,

. . . .

$C_{n+1}^{n+1} = 1$ фигуру типа T_n.

Упражнение. Нарисуйте схематически фигуру типа T_5 и определите число фигур типа T_i, в ней содержащихся ($i = 0, 1, \ldots, 5$).

ГЛАВА V

Топология

Введение

В середине XIX столетия возникло совершенно новое течение в геометрии, которому было суждено вслед за тем стать одной из главных движущих сил современной математики. Предметом новой отрасли, называемой топологией (или analysis situs), является изучение свойств геометрических фигур, сохраняющихся даже тогда, когда эти фигуры подвергаются таким преобразованиям, которые уничтожают все их и метрические, и проективные свойства.

Одним из великих геометров этой эпохи был А. Ф. Мёбиус (1790–1868), человек, не слишком преуспевший из-за своей чрезмерной скромности в научной карьере: он занимал должность астронома в одной из второразрядных немецких обсерваторий. В возрасте шестидесяти восьми лет он представил Парижской Академии мемуар об «односторонних» поверхностях, содержащий кое-какие из наиболее изумительных фактов в новой отрасли геометрии. Подобно многим другим важным научным работам, его рукопись несколько лет валялась на полках Академии, пока обстоятельства не сложились так, что ее опубликовал сам автор. Независимо от Мёбиуса гёттингенский астроном И. Листинг (1808–1882) сделал подобные же открытия и, под влиянием Гаусса, в 1847 г. издал небольшую книгу «Vorstudien zur Topologie». Когда Бернгард Риман (1826–1866) прибыл в Геттинген, чтобы стать там студентом, математическая атмосфера этого университетского города уже была насыщена острым любопытством по отношению к новым и странным геометрическим идеям. Скоро он осознал, что именно в них нужно искать разгадку самых глубоких свойств аналитических функций комплексного переменного. Позднейшее развитие топологии, вероятно, едва ли обязано чему-либо в такой степени, как великолепному зданию римановой теории функций, в которой топологические концепции имеют самое фундаментальное значение.

На первых порах своеобразие методов, которыми приходилось действовать в новой области, воспрепятствовало тому, чтобы полученные здесь результаты были изложены в традиционной дедуктивной форме,

типичной для элементарной геометрии.

Происходило нечто совсем иное: так, Пуанкаре, делая смелые шаги вперед, был вынужден широко и откровенно опираться на геометрическую интуицию. Даже в наши дни изучающий топологию явственно ощущает, что при слишком большой заботе о формальной безупречности существенно геометрическое содержание упускается из виду и тонет в массе деталей. Впрочем, как бы то ни было, нужно рассматривать как особое достижение то обстоятельство, что самые недавние работы по топологии включили эту отрасль геометрии в круг вполне строго построенных математических дисциплин, для которых интуиция была и остается источником, но не конечным критерием истины. По мере развития процесса «формализации» топологии, идущего от Л. Э. Я. Брауэра, удельный вес топологии по отношению к математике в целом непрерывно возрастал. Существенные успехи в указанном направлении принадлежат американским математикам, в частности, О. Веблену, Дж. У. Александеру и С. Лефшетцу.

Хотя топологию можно с полной определенностью назвать продуктом последнего столетия, необходимо все же отметить, что еще и раньше было сделано несколько открытий, которые, как вытекает из современной систематики математических знаний, имеют ближайшее отношение к топологии. Из них самым крупным, несомненно, является установление формулы, связывающей числа вершин, ребер и граней простого многогранника: она была подмечена уже Декартом в 1640 г., позднее переоткрыта и использована Эйлером в 1752 г.; характерные черты топологического утверждения в этой формуле стали очевидными гораздо позднее — после того как Пуанкаре в «формуле Эйлера» и ее обобщениях усмотрел одну из центральных теорем топологии. Итак, по причинам как исторического, так и внутреннего порядка мы начнем наше знакомство с топологией именно с формулы Эйлера. Так как при первых шагах в неизведанной области идеал безупречной строгости вовсе не обязателен и даже мало желателен, то мы будем иногда без колебаний апеллировть непосредственно к интуиции читателя.

§ 1. Формула Эйлера для многогранников

Хотя в античной геометрии изучение многогранников занимало одно из центральных мест, только Декарту и Эйлеру было суждено открыть следующее предложение: *пусть V — число вершин простого многогранника, E — число ребер, F — число граней: тогда*

$$V - E + F = 2. \tag{1}$$

Под *многогранником* здесь подразумевается тело, поверхность которого состоит из конечного числа граней, имеющих форму многоугольников. В случае *правильных* многогранников все многоугольники кон-

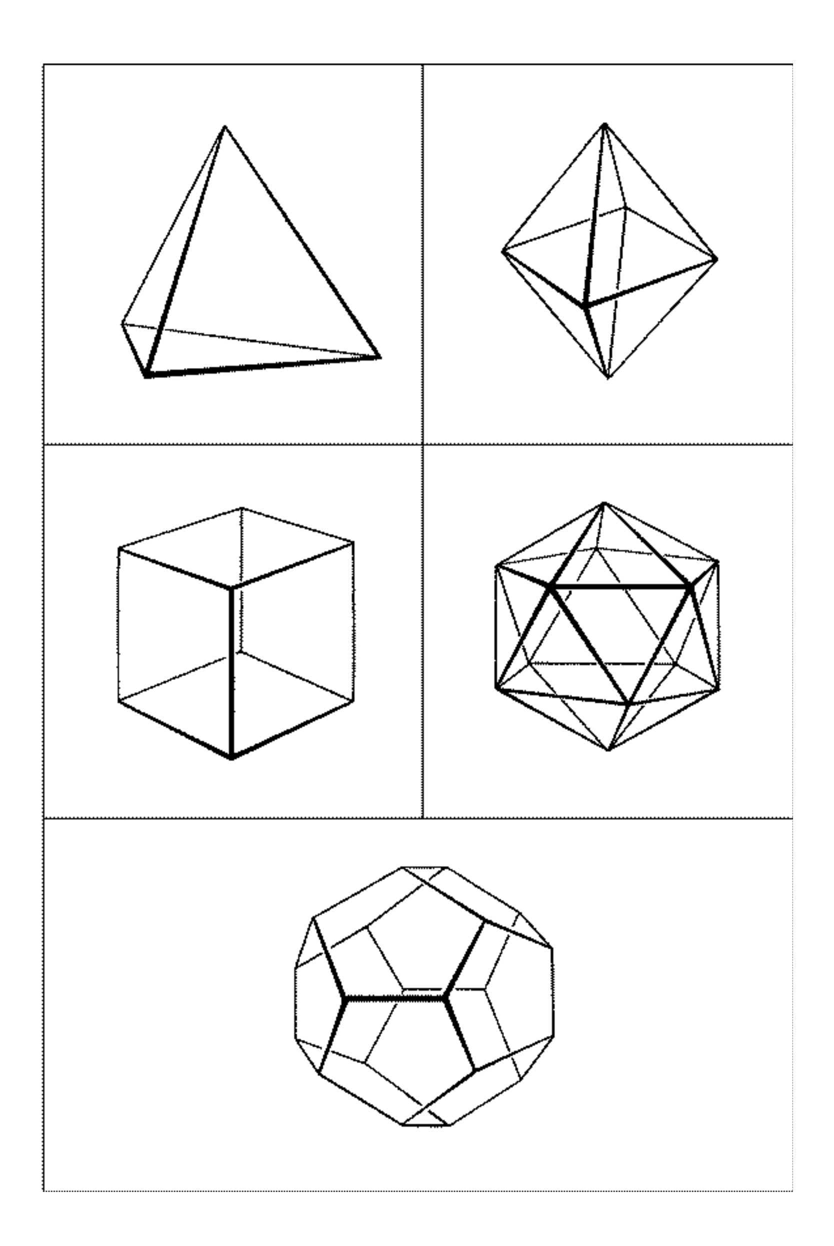


Рис. 119. Правильные многогранники

груэнтны и все плоские углы при вершинах равны между собой. Многогранник называется *простым*, если в нем нет «дыр», так что посредством непрерывной деформации его поверхность может быть переведена в поверхность сферы. На рис. 120 изображен простой многогранник, который не является правильным; на рис. 121 изображен многогранник, не являющийся простым.

Предлагаем читателю проверить справедливость формулы Эйлера для всех многогранников, представленных на рис. 119 и 120; но пусть он убедится также, что для многогранника на рис. 121 эта формула неверна.

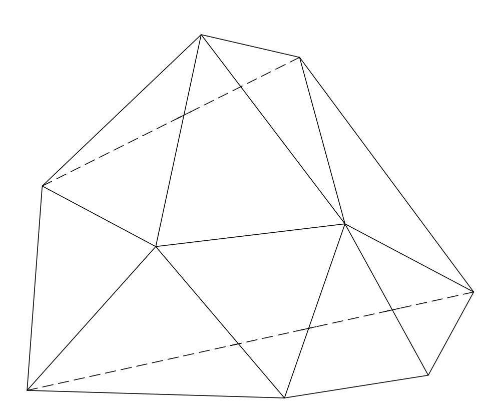


Рис. 120. Простой многогранник: $V - E + F = 9 - 18 + 11 = 2$

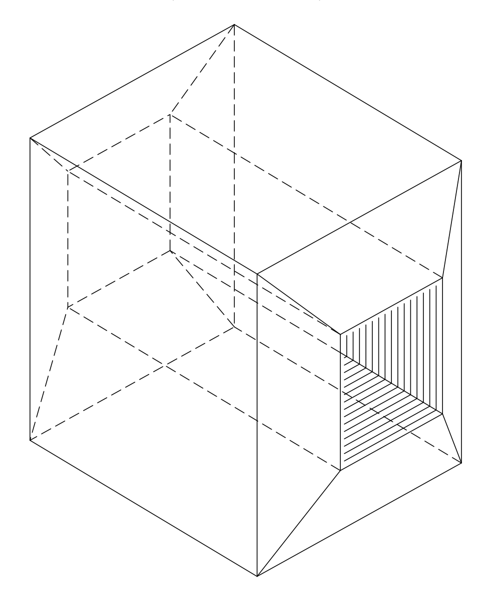


Рис. 121. Непростой многогранник: $V - E + F = 16 - 32 + 16 = 0$

Переходя к доказательству формулы Эйлера, вообразим, что наш многогранник — внутри пустой и что поверхность его сделана из тонкой резины. Тогда, вырезав предварительно одну из граней пустого внутри многогранника, можно оставшуюся поверхность деформировать таким образом, что она расстелется по плоскости. Конечно, при этом и грани многогранника и углы между ребрами испытают резкие изменения. Но «сетка», составленная из вершин и ребер на плоскости, будет содержать то же число вершин и ребер, что и первоначальный многогранник, тогда как число граней станет на одну меньше, так как одна грань была вырезана. Мы убедимся теперь, что для полученной нами сетки на плоскости будет справедливо равенство $V - E + F = 1$; тогда, добавляя вырезанную грань, для первоначального многогранника получим равенство $V - E + F = 2$.

Прежде всего «триангулиру-

ем» плоскую сетку следующим образом. Если в сетке имеются многоугольники с числом углов большим трех, то, выбрав один из них, проведем в нем какую-нибудь диагональ. В результате каждое из чисел E и F увеличится на единицу, но значение выражения $V - E + F$ от этого не изменится. Будем и дальше проводить диагонали, соединяя пары точек (рис. 122), пока сетка не окажется состоящей из одних только треугольников (в чем и заключается наша ближайшая цель). В триангулированной сетке величина $V - E + F$ имеет то же значение, какое имела и до триангуляции, так как проведение каждой новой диагонали

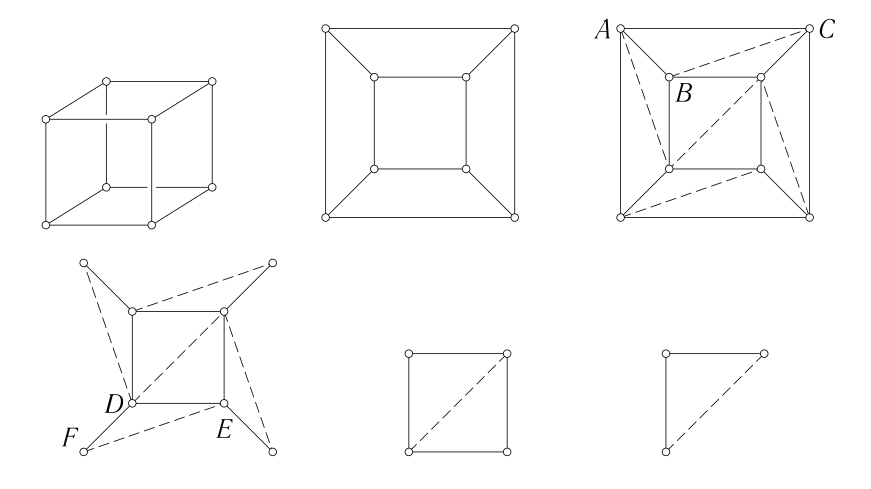


Рис. 122. Доказательсиво теоремы Эйлера

этого значения не меняет. Некоторые из треугольников, далее, имеют ребра (проще сказать — стороны), принадлежащие к «границе» триангулированной сетки. Некоторые из этих треугольников (например, ABC) имеют лишь одно ребро на границе, другие — по два. Возьмем один из такого рода «граничных» треугольников и удалим из него все то, что не принадлежит какому-нибудь другому треугольнику. Так, в треугольнике ABC удалим ребро AC и саму грань, оставляя вершины A, B, C и ребра AB и BC, но в треугольнике DEF удалим грань, два ребра DF и FE и вершину F. При «уничтожении» треугольника ABC числа E и F уменьшаются на 1, а V не изменяется, так что $V - E + F$ также не изменяется. При уничтожении треугольника типа DEF число V уменьшится на 1, E на 2 и F на 1, так что опять-таки $V - E + F$ не изменится. Последовательное осуществление таких удалений граничных треугольников (причем всякий раз меняется и сама граница) приводит, наконец, к одному-единственному треугольнику, имеющему, очевидно, три ребра, три вершины и одну грань. Для образуемой им совсем простой

сетки $V - E + F = 3 - 3 + 1 = 1$. Но мы видели, что при удалении из сетки каждого треугольника $V - E + F$ не изменялось. Значит, $V - E + F$ должно было равняться единице и для первоначальной плоской сетки, а также и для того многогранника с вырезанной гранью, из которого была получена плоская сетка. Отсюда следует, что для исходного многогранника (до вырезания грани) должно было иметь место равенство $V - E + F = 2$. Этим и заканчивается доказательство теоремы Эйлера.

С помощью теоремы Эйлера легко показать, что существует не более пяти типов правильных многогранников. Предположим, что правильный многогранник имеет F граней, из которых каждая есть правильный n-угольник, и что у каждой вершины сходится r ребер. Считая ребра один раз по граням, другой — по вершинам, получим, во-первых,

$$nF = 2E \tag{2}$$

(так как каждое ребро принадлежит двум граням и, следовательно, считается дважды в произведении nF), и, во-вторых,

$$rV = 2E \tag{3}$$

(так как каждое ребро упирается в две вершины). Тогда равенство Эйлера (1) нам дает

$$\frac{2E}{n} + \frac{2E}{r} - E = 2,$$

или

$$\frac{1}{n} + \frac{1}{r} = \frac{1}{2} + \frac{1}{E}. \tag{4}$$

Заметим прежде всего, обращаясь к рассмотрению последнего соотношения, что $n \geqslant 3$ и $r \geqslant 3$, так как многоугольник имеет не меньше трех сторон и в каждой вершине сходится не менее трех граней. С другой стороны, оба числа n и r не могут быть *более* 3, так как в противном случае левая часть равенства (4) не превышала бы $\frac{1}{2}$ и равенство было бы невозможно ни при каком положительном значении E. Итак, нам остается выяснить, какие значения может принять r, если $n = 3$, и какие значения может принять n, если $r = 3$. Подсчитав все возникающие возможности, мы получим число типов правильных многогранников.

При $n = 3$ равенство (4) принимает вид

$$\frac{1}{r} - \frac{1}{6} = \frac{1}{E};$$

r может здесь равняться 3, 4 или 5 (6 или большее значение исключается, так как $\frac{1}{E}$ положительно). При этих значениях n и r оказывается, что E соответственно равно 6, 12 или 30. Так получаются многогранники: тетраэдр, октаэдр и икосаэдр.

Таким же образом при $r = 3$ равенство (4) принимает вид

$$\frac{1}{n} - \frac{1}{6} = \frac{1}{E},$$

из которого следует, что $n = 3$, 4 или 5 и, соответственно $E = 6$, 12 или 30. Получаются многогранники: тетраэдр, куб и додекаэдр.

Подставляя полученные значения n, r и E в соотношения (2) и (3), мы установим число вершин V и число граней F соответствующих многогранников.

§ 2. Топологические свойства фигур

1. Топологические свойства. Мы установили, что формула Эйлера справедлива для случая любого простого многогранника. Но эта формула не теряет смысла и значимости также и применительно к иным, гораздо более общим случаям: вместо многогранников элементарной геометрии с плоскими гранями и прямыми ребрами можно взять простые «многогранники», у которых «гранями» будут кривые поверхности, а «ребрами» — кривые линии, или можно нарисовать «грани» и «ребра» на поверхности, например, шара. Больше того, вообразим, что поверхность многогранника или сферы сделана из тонкого слоя резины; тогда формула Эйлера сохранится, как бы ни была деформирована рассматриваемая поверхность — путем изгибаний, сжатий, растяжений и т. д.,— лишь бы резиновый слой не был порван. Действительно, формула Эйлера относится только к *числу* вершин, ребер и граней; длины же, площади, двойные отношения, кривизна и т. п., как и иные понятия элементарной или проективной геометрии, в данном случае никакой роли не играют.

Мы уже указывали, что элементарная геометрия имеет дело с величинами (расстояния, углы, площади), которые не меняют своих значений при движениях рассматриваемых фигур, тогда как проективная геометрия занимается такими понятиями (точка, прямая, отношение инцидентности, двойное отношение), которые сохраняются при более широкой группе проективных преобразований. Однако и движения, и проективные преобразования — только очень частные случаи гораздо более общих *топологических преобразований*; топологическое преобразование одной геометрической фигуры A в другую A' определяется как произвольное соответствие $p \longleftrightarrow p'$ между точками p фигуры A и точками p' фигуры A', обладающее следующими свойствами:

1. Взаимной однозначностью. (Это значит, что каждой точке p фигуры A сопоставлена одна и только одна точка p' фигуры A', и обратно.)

2. Взаимной непрерывностью. (Это значит, что если мы возьмем две точки p, q фигуры A и станем двигать p так, чтобы расстояние между p и q неограниченно уменьшалось, то расстояние между соответствующими точками p' и q' фигуры A' также будет неограниченно уменьшаться, и обратно.)

Всякое свойство геометрической фигуры A, которое сохраняется также и для той фигуры A', в которую A переходит при топологическом преобразовании, называется *топологическим свойством* фигуры A; *топология* же — это та отрасль геометрии, которая рассматривает исклю-

чительно топологические свойства фигур. Представьте себе, что некоторая фигура должна быть скопирована от руки совершенно малоопытным, но очень добросовестным чертежником, который невольно искривляет прямые линии, искажает углы, расстояния и площади; тогда на сделанной им копии, хотя метрические и проективные свойства фигуры, может быть, и не сохранятся, но топологические свойства все же останутся в неприкосновенности.

Наиболее наглядными примерами топологических преобразований могут служить *деформации*. Вообразите, что фигура вроде сферы или треугольника сделана из тонкого слоя резины (или нарисована на таковом), и затем растягивайте и крутите резину самыми разнообразными способами, лишь бы не рвать ее и не приводить двух различных точек в

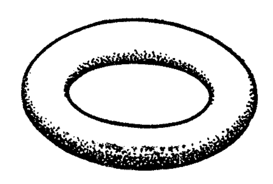
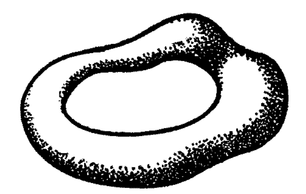



Рис. 123. Поверхности, топологически эквивалентные

состояние физического совпадения. (Приведение двух различных точек в состояние физического совпадения нарушило бы условие 1. Разрыв резинового слоя противоречил бы условию 2: действительно, рассматривая две точки, лежащие по разные стороны линии разрыва, мы видим, что расстояние между ними может быть неограниченно малым, тогда как после разрыва этого уже не будет.)

Фигура в окончательном ее положении — после указанных операций — будет находиться в топологическом соответствии с фигурой в ее первоначальном положении. Треугольник можно деформировать в другой треугольник, или в окружность, или в эллипс, и потому названные фигуры обладают совершенно одинаковыми топологическими свойствами. Но никак нельзя деформировать круг в отрезок прямой или поверхность сферы в боковую поверхность цилиндра.

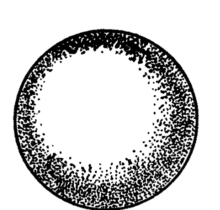
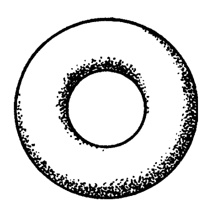



Рис. 124. Поверхности, топологически неэквивалентные

Но общее понятие топологического преобразования шире, чем понятие деформации. Например, если фигура разрезана до деформации и склеена по тем же

линиям после деформации, то в итоге, несомненно, получается некоторое топологическое преобразование первоначальной фигуры, хотя это преобразование может и не быть деформацией. Так, две кривые, изображенные на рис. 134 (стр. 275), топологически эквивалентны друг другу и эквивалентны каждая окружности, так как их можно разрезать, распутать и снова склеить. Но предварительно не разрезав, невозможно одну кривую деформировать в другую.

Топологические свойства фигур (вроде того свойства, которое дается теоремой Эйлера, или других, которые будут рассмотрены ниже) представляют величайший интерес во многих математических исследованиях. В известном смысле это — самые глубокие, самые основные геометрические свойства, так как они сохраняются при самых «резких» преобразованиях.

2. Свойства связности. В качестве следующего примера фигур, топологически неэквивалентных, рассмотрим две плоские области на рис. 125. Первая состоит из всех внутренних точек круга; вторая — из всех точек, расположенных между двумя концентрическими кругами. Любая замкнутая кривая, лежащая в области *а*, может быть непрерывно деформирована, или «сжата», в одну точку, *не выходя из этой области.* Область, обладающая таким свойством, называется *односвязной.* Что касается области *б*, то она не односвязна. Так, окружность, концентрическая с двумя граничными окружностями и лежащая между ними, не может быть сжата в точку, не выходя из области, так как во время деформации кривая должна будет пройти через общий центр кругов, а он не принадлежит рассматриваемой области. Область, которая не является односвязной, называется *многосвязной.* Если *двусвязную* область разрезать вдоль одного из радиусов, как это сделано на рис. 126, то полученная область становится односвязной.

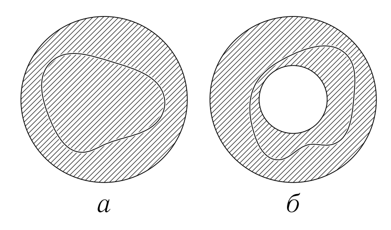


Рис. 125. Односвязная и двусвязная области

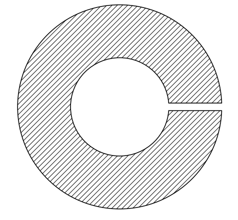


Рис. 126. После разреза двусвязная область становится односвязной

Вообще, можно построить области с двумя, тремя или большим количеством «дыр». Область с двумя «дырами» изображена на рис. 127; чтобы превратить ее в односвязную, нужно сделать два разреза. Если

нужно сделать $n-1$ взаимно не пересекающихся разрезов от границы к границе, чтобы превратить данную многосвязную область в односвязную, то говорят, что область имеет *порядок связности* n. Порядок связности плоской области представляет собой важный топологический инвариант этой области.

§ 3. Другие примеры топологических теорем

1. Теорема Жордана о замкнутой кривой. На плоскости нарисована простая замкнутая кривая (нигде сама себя не пересекающая). Посмотрим, какое свойство этой фигуры сохраняется неизменным даже в том случае, если плоскость будет подвергаться каким угодно деформациям, как будто бы она была сделана из тонкого слоя резины. Длина кривой или площадь ограниченной ею части плоскости при деформациях не сохраняется. Но у рассматриваемой конфигурации есть и топологическое свойство, столь простое, что может показаться тривиальным. *Простая замкнутая кривая C на плоскости делит плоскость ровно на две области, внутреннюю и внешнюю.* Точнее говоря, мы утверждаем следующее: точки плоскости разбиваются на два класса — A (внешние точки) и B (внутренние точки) — таким образом, что любая пара точек, принадлежащих одному и тому же классу, может быть связана кривой, не имеющей общих точек с C, тогда как всякая кривая, соединяющая две какие-нибудь точки разных классов, непременно пересекается с C. Это утверждение вполне очевидно, например, для случая окружности или эллипса, но уже чуть менее очевидно для такой сложной кривой, как причудливой формы многоугольник, изображенный на рис. 128.

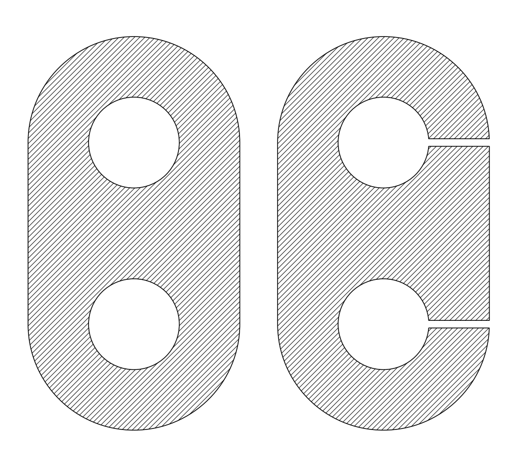


Рис. 127. Редукция трехсвязной области

Впервые эта теорема была сформулирована Камиллом Жорданом (1838–1922) в его широко известном «Cours d'analyse», из которого целое поколение математиков почерпнуло современную концепцию математической строгости. Как это ни странно, доказательство, данное самим Жорданом, не было ни кратким, ни простым по своей идее, но в особенности удивительно то, что, как оказалось, оно и не было

вполне исчерпывающим, и понадобились значительные усилия, чтобы восполнить его пробелы. Первые строгие доказательства теоремы Жордана были очень сложными и трудно воспринимались даже людьми с хорошей математической подготовкой. Сравнительно простые доказательства были придуманы лишь недавно. Одно из затруднений заключается в большой общности понятия «простой замкнутой» кривой, значительно более широкого, чем понятие многоугольника или «гладкой» кривой: по определению, «простая замкнутая кривая» есть любая кривая, топологически эквивалентная окружности. С другой стороны, необходимо таким терминам, как «внутри» или «вне» (столь ясным интуитивно), дать логические определения, прежде чем строгое доказательство станет возможным. Проанализировать в их полной общности возникающие в связи с этим отношения и концепции есть теоретическая задача первостепенного значения, разрешению которой в большой степени служит современная топология. Но, с другой стороны, следует иметь в виду и то обстоятельство, что, занимаясь изучением конкретных явлений в области геометрии, в громадном большинстве случаев малоуместно вводить понятия, неограниченная общность которых создает излишние затруднения. Так, возвращаясь к теореме Жордана, существенно то, что для случая «хорошо ведущих себя» кривых — например, для многоугольников или для кривых с непрерывно меняющейся касательной (которые только и встречаются в наиболее важных задачах) — доказательство этой теоремы может быть проведено совсем просто. Для случая многоугольников мы укажем доказательство в дополнении к этой главе.

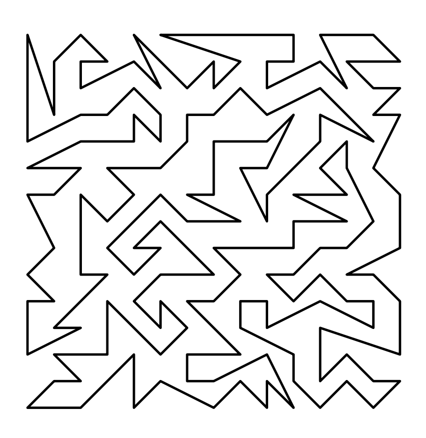


Рис. 128. Какие точки находятся внутри этого многоугольника?

2. Проблема четырех красок. Пример только что рассмотренной теоремы Жордана способен, пожалуй, навести на мысль, что топология занимается придумыванием строгих доказательств для таких истин, в которых не станет сомневаться ни один здравомыслящий человек. Но это совсем не так: существует много вопросов топологического характера, в числе которых иные формулируются чрезвычайно просто и на которые интуиция не дает удовлетворительных ответов. Примером может служить знаменитая «проблема четырех красок».

Раскрашивая географическую карту, обыкновенно стараются рас-

пределить цвета между странами таким образом, чтобы две страны, имеющие общую границу, были окрашены по-разному. Было обнаружено на опыте, что любая карта, сколько бы ни было изображено на ней стран и как бы они ни были расположены, может быть раскрашена с соблюдением указанного правила не более чем *четырьмя* красками. Легко убедиться, что меньшее число достаточным для всех случаев не является. На рис. 129 изображен остров посреди моря, который никак нельзя раскрасить менее чем четырьмя красками, так как на нем имеется четыре страны, из которых каждая соприкасается с остальными тремя.

Тот факт, что до настоящего времени не было найдено такой карты, для раскрашивания которой потребовалось бы более четырех красок, приводит к мысли о справедливости такой теоремы: *при любом данном разбиении плоскости на области, не покрывающие друг друга ни полностью, ни частично, всегда возможно пометить их цифрами* 1, 2, 3, 4 *таким образом, чтобы «прилежащие» области были обозначены разными цифрами.* Под «прилежащими» областями понимаются такие, которые имеют целый отрезок границы общим: две области, имеющие лишь одну общую точку (или даже конечное число общих точек) — как, например, штаты Колорадо и Аризона,— не будут называться «прилежащими», так как никакого смешения или неудобства не возникает, если их раскрасить одинаково.

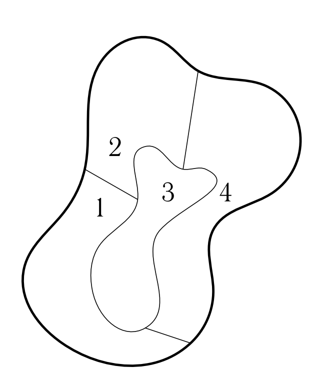


Рис. 129. Раскрашивание карты

Есть основания полагать, что впервые проблема четырех красок была поставлена Мёбиусом в 1840 г.; позднее ее формулировали де Морган в 1850 г. и Кэли в 1878 г. «Доказательство» ее было опубликовано в 1879 г. Кемпе, но Хивуд в 1890 г. нашел ошибку в рассуждении Кемпе. Пересматривая доказательство Кемпе, Хивуд обнаружил, что *пяти* красок всегда достаточно. (Доказательство теоремы о пяти красках дано в приложении к этой главе.) Несмотря на усилия многих выдающихся математиков, положение вплоть до нашего времени остается в сущности неизменным. Было *доказано*, что пяти красок достаточно для всех карт, и *имеется предположение*, что достаточно также четырех. Но, как и в случае знаменитой теоремы Ферма (см. стр. 60), ни доказательства этого предположения, ни противоречащего ему примера приведено не было, и указанное предположение остается одной из нере-

шенных «больших» математических проблем[1]. Заметим, между прочим, что проблема четырех красок была решена в положительном смысле для частных случаев, когда число областей не превышает *тридцати восьми*. Отсюда ясно, что если в общем случае теорема неверна, то опровергающий пример должен быть не особенно простым.

В рассматриваемой проблеме четырех красок предполагается, что карта нарисована или на плоскости, или на сфере. Эти два случая эквивалентны. В самом деле, каждая карта, заданная на сфере, может быть перенесена па плоскость, если проделаем дырочку внутри одной из областей A и затем расплющим оставшуюся часть сферы по плоскости, как мы это делали при доказательстве теоремы Эйлера. Полученная карта на плоскости покажет нам «остров», состоящий из всех нетронутых областей, и «море», состоящее из одной области A. С другой стороны, проделывая всю эту процедуру в обратном направлении, можно любую карту на плоскости превратить в карту на сфере. Итак, вместо карт на плоскости можно ограничиться рассмотрением карт на сфере. Больше того, так как деформации областей и их границ существенно не влияют на нашу проблему, то можно предположить, что граница каждой области есть простой замкнутый многоугольник, состоящий из дуг больших кругов. Но даже таким образом «регуляризированная» проблема не решена; трудности в данном случае (не в пример теореме Жордана) зависят не от общности понятия области и кривой.

В связи с проблемой четырех красок стоит отметить то замечательное обстоятельство, что для некоторых поверхностей более сложного типа, чем плоскость или сфера, соответствующие теоремы действительно были доказаны, так что, как это ни парадоксально, анализ более сложных (в геометрическом отношении) поверхностей в данном случае проводится легче, чем более простых. Например, было установлено для случая поверхности тора, имеющей вид «бублика» (см. рис. 123), что всякая нарисованная на ней «карта» может быть раскрашена *семью* красками и что, с другой стороны, на ней мыслимы такие «карты», составленные из семи областей, что каждая область соприкасается с остальными шестью.

***3. Понятие размерности.** Понятие о «числе измерений», или о «размерности», не представляет особых затруднений, пока речь идет о таких простых геометрических образах, как точки, линии, треугольники или многогранники. Отдельная точка или любое *конечное* множество точек имеет размерность нуль, отрезок — размерность 1, поверхность треугольника или сферы —

[1] Проблема четырех красок была решена в 1976 г. Ее решение свелось к проверке 1482 карт и перебору различных комбинаций раскрасок каждой из них. Перебор был осуществлен с помощью компьютера; многие математики полагают, что рассуждение, опирающееся на компьютерный перебор, нельзя считать убедительным. — *Прим. ред. наст. изд.*

размерность 2. Множество всех точек куба имеет размерность 3. Однако при желании обобщить понятие размерности на точечные множества более общих типов возникает необходимость в точном определении. Какую размерность следует, например, приписать множеству R, состоящему из всех точек прямой, у которых координаты — *рациональные* числа? Множество рациональных точек на прямой всюду плотно, и потому, казалось бы, ему, как и самому отрезку прямой, надлежало бы приписать размерность 1. С другой стороны, между всякими двумя рациональными точками существуют иррациональные «дыры», как между всякими двумя точками конечного множества, и это говорит в пользу размерности 0.

Еще запутаннее обстоит дело с размерностью любопытного множества, впервые рассмотренного Кантором, построенного следующим образом. Из единичного отрезка $0 \leqslant x \leqslant 1$ удалим среднюю треть (интервал), т. е. все точки x, удовлетворяющие неравенству $\frac{1}{3} < x < \frac{2}{3}$. Оставшееся точечное множество обозначим через C_1. Множество C_1 состоит из двух отрезков; удалим теперь из каждого отрезка его среднюю треть, и то множество, которое останется, обозначим через C_2. Повторим опять эту процедуру, удаляя среднюю треть у всех четырех отрезков; получим C_3. Дальше таким же образом получим C_4, C_5, C_6, ... Обозначим через C множество точек, которое останется, когда все средние трети будут удалены; другими словами, C есть множество точек, принадлежащих одновременно всем множествам C_1, C_2, C_3, ... В первой операции был удален интервал длины $\frac{1}{3}$; во второй операции — два интервала, каждый длины $\frac{1}{9}$ и т. д.; сумма длин всех удаленных интервалов равна

$$1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3^2} + 2^2 \cdot \frac{1}{3^3} + \ldots = \frac{1}{3}\Big(1 + \frac{2}{3} + \Big(\frac{2}{3}\Big)^2 + \ldots\Big).$$

Бесконечный ряд в больших скобках есть геометрическая прогрессия, сумма которой равна $\frac{1}{1 - \frac{2}{3}} = 3$; итак, сумма длин удаленных промежутков составляет 1. И все-таки далеко не все точки отрезка удалены: множество C не пустое. Например, все точки, являющиеся концами удаленных интервалов —

$$\frac{1}{3}, \quad \frac{2}{3}; \quad \frac{1}{9}, \quad \frac{2}{9}, \quad \frac{7}{9}, \quad \frac{8}{9}; \quad \ldots$$

— ему принадлежат. Можно легко убедиться, что множество C состоит в точности из всех тех чисел x, разложения которых в бесконечную дробь по основанию 3 могут быть написаны в форме

$$x = \frac{a_1}{3} + \frac{a_2}{3^2} + \frac{a_3}{3^3} + \ldots + \frac{a_n}{3^n} + \ldots,$$

где всякое a_n есть 0 или 2 тогда как в аналогичном разложении для всякой удаленной точки среди чисел a_n, хоть раз встретится 1.

Какова же размерность множества C? Диагональный процесс, с помощью которого была доказана несчетность множества всех действительных чисел, может быть видоизменен таким образом, чтобы тот же результат получился и для множества C. Отсюда было бы естественно заключить, что множеству C надлежит приписать размерность 1. С другой стороны, C не содержит никакого, даже самого малого, промежутка, как и любое конечное множество;

это сближает C с множествами размерности 0. Таким же образом, восставив в *плоскости* x, y из каждой рациональной точки или из каждой точки канторова множества перпендикуляр длины 1 к оси x (направляя его в сторону положительных значений y), мы получим множества, относительно которых может возникнуть сомнение — приписать ли ему размерность 2 или 1.

Впервые Пуанкаре (в 1912 г.) обратил внимание на необходимость более глубокого анализа и более точного определения размерности. Пуанкаре заметил, что прямая или кривая имеет размерность 1, так как любые две точки на ней можно разделить, удаляя одну-единственную точку (множество размерности 0); плоскость же имеет размерность 2 по той причине, что для разделения двух точек на плоскости нужно удалить целую замкнутую кривую (множество размерности 1). Это приводит к мысли о том, что понятие размерности имеет «индуктивную» природу: некоторому «пространству» следует приписать размерность n, если две точки в нем разделяются при удалении подмножества точек размерности $n-1$ (но удаления подмножества меньшей размерности уже не было бы достаточно). В сущности, такого рода индуктивное определение неявно содержится уже в евклидовых «Началах», где одномерный образ толкуется как нечто, граница чего состоит из точек; двумерный образ — как нечто, граница чего состоит из линий; наконец, трехмерный образ — как нечто, граница чего состоит из поверхностей.

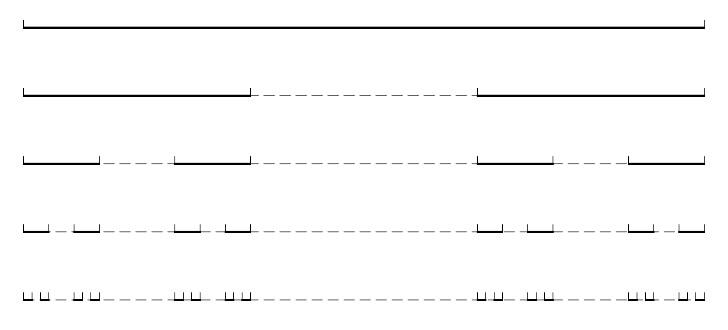


Рис. 130. Канторово множество

За последние годы была развита обширная теория — теория размерности. Определение размерности начинается с того, что разъясняется смысл термина «точечное множество размерности 0». Любое *конечное* точечное множество обладает тем свойством, что каждая его точка может быть заключена в сколь угодно малую область пространства, причем на границе области нет точек множества. Это свойство принимается теперь за определение размерности 0. Условимся ради удобства говорить, что пустое множество имеет размерность -1. В таком случае множество S имеет размерность 0, если оно не имеет размерности -1 (т. е. если S содержит хоть одну точку) и если каждая точка S может быть заключена в произвольно малую область, граница которой пересекает S по множеству размерности -1 (т. е. совсем не содержит ни одной точки S). Так, например, множество рациональных точек на прямой имеет размерность 0, так как каждая рациональная точка может быть рас-

сматриваема как центр произвольно малого промежутка с иррациональными концами. Канторово множество C также размерности 0, так как, подобно множеству рациональных точек, оно получается посредством удаления везде плотного множества точек прямой.

Итак, мы уже определили понятия «размерность -1» и «размерность 0». Теперь легко понять, что такое «размерность 1»: говорят, что множество S имеет «размерность 1», если оно не есть ни размерности -1, ни размерности 0, и если каждая точка S может быть заключена в произвольно малую область, граница которой пересекается с S по множеству размерности 0. Отрезок прямой обладает этим свойством, так как границей каждого промежутка является пара точек, т. е. множество размерности 0 по предыдущему определению. Дальше, продолжая таким же образом, мы можем последовательно определить, что такое размерность 2, размерность 3 и т. д., причем каждое следующее определение основывается на предыдущем.

Таким образом, говорят, что множество S имеет размерность n, если оно не имеет меньшей размерности и если каждая точка S может быть заключена в произвольно малую область, граница которой пересекается с S по множеству размерности $n-1$. Например, плоскость имеет размерность 2, так как любая точка плоскости может быть заключена в кружок произвольно малого радиуса, граница которой имеет размерность 1.[1] В обыкновенном пространстве никакое множество точек не может иметь размерность большую чем 3, так как любая точка пространства есть центр произвольно малой сферы, граница которой имеет размерность 2. Но в современной математике термин «пространство» употребляется в более общем смысле; он обозначает любую систему объектов, для которой введено понятие «расстояния» или «окрестности», и такого рода абстрактные «пространства» могут иметь размерность большую чем 3. Простым примером является *декартово n-мерное пространство*, «точки» которого суть системы из n действительных чисел, взятых в определенном порядке:

$$P = (x_1, x_2, \ldots, x_n),$$
$$Q = (y_1, y_2, \ldots, y_n),$$

а «расстояние» между P и Q определяется по формуле

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}.$$

Можно показать, что это пространство имеет размерность n. Пространство, которое не имеет размерности n, как бы велико ни было n, называется пространством бесконечной размерности. Известно много примеров таких пространств.

В теории размерности устанавливается одно чрезвычайно интересное свойство двумерных, трехмерных и вообще n-мерных фигур. Начнем с двумер-

[1] Сказанное не означает, что доказательство того, что плоскость имеет размерность 2 в смысле нашего определения, уже закончено: остается доказать, что граница круга (окружность) имеет размерность 1, и что сама плоскость не имеет размерности 0 или 1. Эти утверждения можно доказать, как и аналогичные утверждения для высших размерностей. Все предыдущие рассуждения показывают, что приведенное выше общее определение размерности не стоит в противоречии с обычным его пониманием.

ного случая. Если какая-то простая двумерная фигура подразделена на достаточно маленькие «ячейки» (причем предполагается, что каждая ячейка содержит свою границу), то непременно найдутся такие точки, которые принадлежат сразу *по меньшей мере трем ячейкам, какова бы ни была форма выбранных ячеек.* Вместе с тем *существуют такие* разбиения фигуры на ячейки, что никакая точка фигуры не принадлежит сразу *больше чем трем* ячейкам. Так, если рассматриваемая двумерная фигура есть квадрат (рис. 131), то непременно имеются точки вроде той, которая сразу принадлежит трем ячейкам 1, 2 и 3, но для указанного на рисунке разбиения не существует точки, которая сразу принадлежала бы большему числу ячеек. Точно так же в трехмерном случае можно доказать, что если некоторая объемная фигура (тело) разбита на достаточно маленькие ячейки, то наверняка существуют точки, принадлежащие по меньшей мере четырем ячейкам, и вместе с тем можно выбрать такие подразделения, что никакая точка не будет принадлежать сразу больше чем четырем ячейкам.

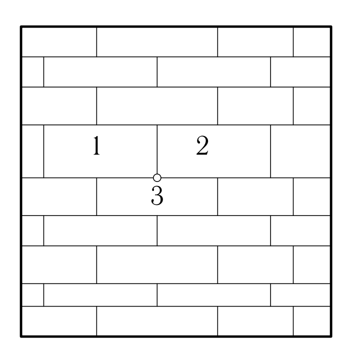


Рис. 131. Теорема о покрытии

Все эти соображения приводят нас к следующей теореме, высказанной А. Лебегом и Брауэром: *если n-мерная фигура разбита на достаточно маленькие ячейки, то непременно существуют точки этой фигуры, принадлежащие сразу по меньшей мере $n+1$ ячейкам; вместе с тем возможно указать и такие разбиения, что ни одна точка фигуры не будет принадлежать сразу более чем $n+1$ ячейкам.* Эта теорема характеризует размерность рассматриваемой фигуры: все фигуры, для которых теорема верна, являются n-мерными, все прочие имеют иную размерность. По этой причине указанная теорема может быть взята за определение размерности (так и делают некоторые авторы).

Размерность фигуры относится к числу топологических ее свойств: никакие две фигуры различных размерностей не могут быть топологически эквивалентными. В этом заключается замечательная теорема об «инвариантности размерности»: чтобы оценить ее должным образом, стоит напомнить другую теорему (доказанную на стр. 106), согласно которой множество точек квадрата имеет ту же мощность, что и множество точек отрезка. Соответствие между точками, установленное при доказательстве этой теоремы, не топологическое, так как требование непрерывности нарушается.

4. Теорема о неподвижной точке. В приложениях топологии к другим отраслям математики играют важную роль теоремы о «неподвижной точке». Типическим примером является излагаемая ниже теорема Брауэра. Она гораздо менее «очевидна» в интуитивном смысле, чем другие топологические теоремы.

Рассмотрим круглый диск на плоскости. Под таковым мы пони-

маем внутренность некоторого круга вместе с его границей (окружностью). Предположим, что весь этот диск подвергается некоторому топологическому преобразованию (даже не обязательно взаимно однозначному), при котором всякая точка диска остается точкой диска, хотя и меняет свое положение. Например, представляя себе этот диск сделанным из тонкой резины, можно его сжимать, растягивать, вращать, изгибать — одним словом, деформировать как угодно, лишь бы его точки не вышли за пределы первоначального положения диска. Иначе еще можно представить себе, что жидкость, налитая в стакан, приведена в движение таким образом, что частицы, находившиеся на поверхности, остаются на ней и во время движения; тогда в каждый определенный момент времени положение частиц на поверхности определяет некоторое топологическое преобразование или трансформацию первоначального их распределения. Теорема Брауэра утверждает: *каждое непрерывное преобразование такого рода оставляет неподвижной по крайней мере одну точку*; другими словами, существует по меньшей мере одна точка, положение которой после преобразования совпадает с положением ее до преобразования. (В примере с жидкостью неподвижные точки зависят от избранного момента времени; в частности, если движение сводится к простому круговому вращению, то неподвижной точкой в любой момент является центр.) Излагаемое далее доказательство существования неподвижной точки — очень характерный пример рассуждений, применяемых в топологии.

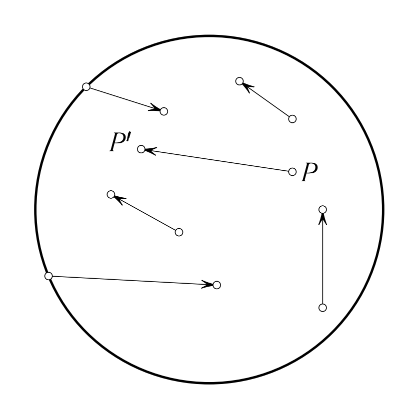


Рис. 132. Векторы преобразования

Рассмотрим наш диск до и после преобразования и допустим, что, вопреки утверждению теоремы, *ни одна* точка не остается неподвижной, так что любая точка диска после преобразования превращается в некоторую *другую* точку диска. Каждой точке P диска в его первоначальном положении сопоставим стрелку или «вектор преобразования» PP', причем P' есть та точка, в которую переходит P после преобразования. Такая стрелка будет выходить из каждой точки диска, так как всякая точка куда-то перемещается. Рассмотрим теперь все точки граничной окружности вместе с соответствующими векторами преобразования. Все эти векторы направлены внутрь круга, так как по предположению ни одна точка не выходит за его пределы. Начнем с какой-нибудь точки P, лежащей на граничной окружности, и пойдем по этой окружности в

направлении, противоположном движению часовой стрелки. При этом направление вектора преобразования будет изменяться, так как различным точкам границы соответствуют различно направленные векторы. Все эти векторы можно также представить себе (подвергнув их параллельному переносу) выходящими из некоторой одной и той же точки плоскости (рис. 133). Легко понять, что, когда мы обойдем один раз весь круг, вектор после ряда поворотов вернется в первоначальное положение. Число полных поворотов, сделанных при этом нашим вектором, мы назовем *индексом* рассматриваемой граничной окружности; точнее говоря, мы определим индекс как *алгебраическую сумму* различных изменений в угле векторов, условливаясь, что всякому частному повороту по часовой стрелке приписывается знак минус, против часовой стрелки — знак плюс. Индекс есть итоговый результат, который a priori равен одному из чисел $0, \pm 1, \pm 2, \pm 3, \ldots$, соответствующих итоговым поворотам на $0°, \pm 360°, \pm 720°, \ldots$ Мы утверждаем теперь, что индекс граничной окружности равен единице, т. е. что итоговый поворот вектора преобразования составляет один полный поворот в положительном направлении. Прежде всего напомним еще раз, что вектор преобразования, имеющий начало в точке граничного круга, направлен непременно внутрь круга, а не по касательной. Если допустить, что итоговый поворот вектора преобразования отличается от итогового поворота *касательного вектора* (а этот последний поворот в точности равен 360°, так как касательный вектор, очевидно, делает один полный поворот), то разность между итоговыми поворотами касательного вектора и вектора преобразования будет равна кратному 360°, но никак не нулю. Отсюда следует, что вектор преобразования при обходе круга должен будет по крайней мере раз сделать полный поворот вокруг касательного вектора,

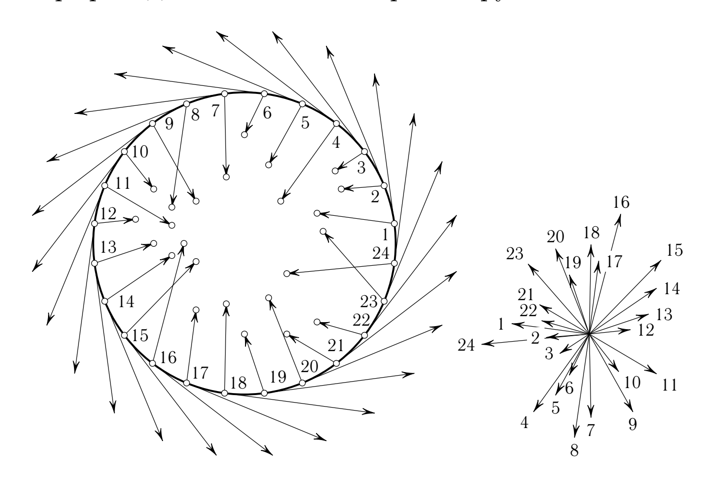


Рис. 133. К доказательству теоремы Брауэра

а так как оба вектора изменяются непрерывно, то в некоторой точке окружности направления двух векторов совпадут. Но это, как мы видели, невозможно.

Рассмотрим теперь окружность, концентрическую границе диска, но с меньшим радиусом, а также соответствующие векторы преобразования. Для этой новой окружности индекс также непременно равен единице. В самом деле, при переходе от граничной окружности к новой окружности индекс должен меняться непрерывно, так как направления самих векторов преобразования меняются непрерывно. Но индекс может принимать только целые значения и потому остается равным единице: действительно, переход от единицы к какому-нибудь другому целому числу обязательно был бы связан со скачком, т. е. нарушением непрерывности. (Очень характерное математическое рассуждение: величина меняется непрерывно, но может принимать только целые значения, значит, она постоянна.) Итак, мы можем найти окружность, концентрическую граничной, притом сколь угодно малую, для которой индекс будет равен единице. Но это невозможно, так как, в силу непрерывности преобразования, векторы преобразования в достаточно малом круге должны весьма мало отличаться от вектора в центре круга. И потому итоговый поворот такого вектора при обходе круга может быть сделан, скажем, меньше 10°, если только радиус круга будет достаточно мал. Но отсюда следует, что индекс такого круга (обязательно целое число) не может быть отличен от нуля. Полученное противоречие показывает, что сделанное нами допущение об отсутствии неподвижных точек преобразования должно быть отвергнуто. Таким образом, теорема доказана.

Теорема о неподвижных точках имеет место не только для кругового диска, но, конечно, и для треугольника, квадрата и всякой другой фигуры, в которую диск может быть переведен топологическим преобразованием. В самом деле, если бы некоторая фигура A, получающаяся из кругового диска посредством такого рода преобразования, могла быть преобразована сама в себя без неподвижных точек, то тем самым было бы определено и топологическое преобразование кругового диска самого в себя без неподвижных точек, а это, как мы видели, невозможно. Теорема обобщается также на случай трехмерных фигур — сфер или кубов, но доказательство не столь просто.

* Хотя теорема Брауэра о неподвижных точках в случае круга не является вполне очевидной в интуитивном смысле, однако легко убедиться, что она является непосредственным следствием такой достаточно очевидной теоремы: *невозможно непрерывно отобразить круговой диск в одну только его граничную окружность таким образом, чтобы каждая точка этой окружности оставалась неподвижной.* Убедимся, что существование непрерывного отображения диска в себя без неподвижных точек противоречит этой последней теореме. Предположим, что указанного рода непрерывное отображение $P \rightarrow$

P' существует. Тогда для всякой точки P нашего диска проведем вектор с началом в точке P', проводя его через P и заканчивая в точке P^*, где он встретится с граничной окружностью. Тогда преобразование $P \to P^*$ будет непрерывным отображением всего диска в граничную окружность, оставляющим неподвижными все точки этой окружности, возможность чего была отвергнута. Подобное рассуждение можно применить и при доказательстве теоремы Брауэра в трехмерном случае сферы или куба.

Легко убедиться, с другой стороны, что для некоторых фигур непрерывные преобразования в себя без неподвижных точек возможны.

Например, кольцеобразная область между двумя концентрическими окружностями может быть подвергнута вращению около центра на угол, не являющийся кратным 360°, и это как раз будет непрерывным преобразованием области в себя без неподвижных точек. Такое же преобразование можно произвести над поверхностью сферы, сопоставляя всякой ее точке диаметрально противоположную. Но, применяя тот же метод, что и в случае диска, не представит труда доказать, что непрерывное преобразование сферической поверхности, не переводящее ни одной точки в диаметрально противоположную (например, всякая малая деформация), непременно имеет неподвижные точки.

Теоремы о неподвижных точках вроде перечисленных выше доставляют могущественный метод для доказательства многих «теорем существования» в разных областях математики, причем геометрический характер этих теорем часто далеко не очевиден. Замечательным примером может служить теорема Пуанкаре, высказанная им незадолго до смерти, в 1912 г., без доказательства. Из этой теоремы непосредственно вытекает существование бесчисленного множества периодических орбит в ограниченной проблеме трех тел. Пуанкаре не сумел обосновать своей догадки, доказательство этого замечательного факта получил через год американский математик Г. Д. Биркгоф. С тех пор топологические методы неоднократно и с большим успехом применялись к изучению качественного поведения динамических систем.

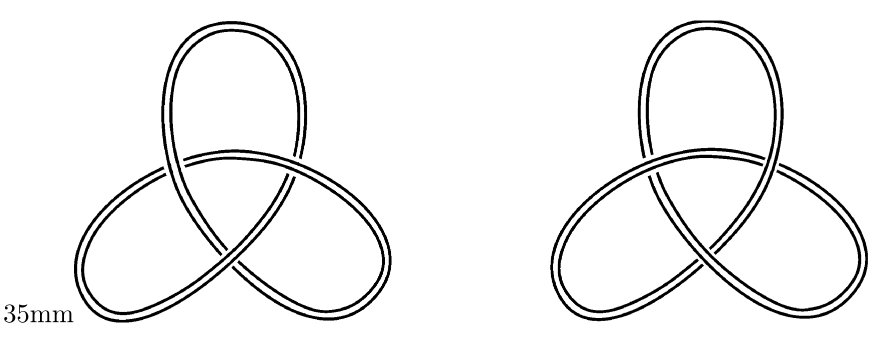


Рис. 134. Топологически эквивалентные узлы, не переводящиеся друг в друга

5. Узлы. В качестве последнего примера отметим, что трудные математические проблемы топологического характера возникают в связи

с изучением узлов. Узел образуется, когда из отрезка веревки делают петли, затем сквозь них пропускают концы веревки и, наконец, два конца соединяют вместе. Изготовленная из веревки замкнутая кривая представляет собой геометрическую фигуру, существенные свойства которой не изменяются, как бы в дальнейшем ни перетягивать или ни перекручивать веревку. Но как возможно было бы дать внутреннюю характеристику, которая позволила бы различить тем или иным способом «заузленные» кривые между собой и отличать их от «незаузленных» вроде круга? Ответ далеко не прост, и еще менее прост исчерпывающий математический анализ узлов разных типов. Затруднения встречаются даже при самых первых шагах в этом направлении. Взгляните на два узла, напоминающие трилистники, изображенные на рис. 134. Они совершенно симметричны друг другу, являются взаимно «зеркальными отображениями», они топологически эквивалентны и вместе с тем неконгруэнтны друг другу. Возникает проблема: можно ли деформировать непрерывным движением один узел в другой? Ответ отрицателен; но доказательство потребовало бы значительно большего владения топологической техникой и бóльших знаний из области теории групп, чем предполагают рамки этой книги.

§ 4. Топологическая классификация поверхностей

1. Род поверхности. Многие простые, но весьма существенные обстоятельства выясняются при изучении двумерных поверхностей. Сравним, например, поверхность сферы с поверхностью тора. Взглянув на рис. 135, сразу можно обнаружить различие: на сфере, как и на плоскости, замкнутая кривая вроде C разделяет поверхность на две части; но на торе существуют и такие замкнутые кривые, например C', которые не разделяют поверхности на две части. Если мы говорим, что кривая C разделяет сферу на две части, то это означает, другими словами, что при разрезании поверхности сферы по кривой C эта поверхность распадается на два не связанных между собой куска, или еще, иначе, что можно найти две такие точки сферы, что всякая кривая на сфере, их соединяющая, непременно пересечется с кривой C. На-

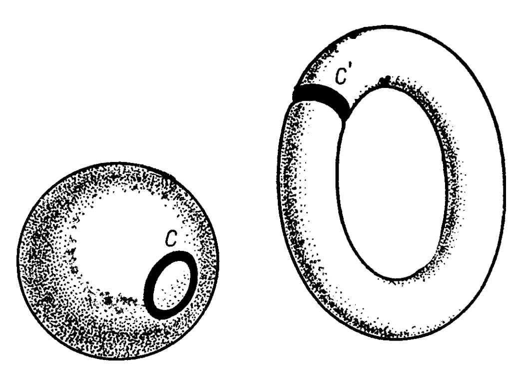


Рис. 135. Разрезы на сфере и на торе

против, если разрезать тор по кривой C', то после разреза поверхность не распадется: любые две ее точки можно соединить кривой, не имеющей общих точек с C'. Указанное различие свидетельствует о том, что сфера и тор в топологическом смысле не принадлежат одному и тому же классу поверхностей: тор нельзя топологически преобразовать в сферу.

Рассмотрим теперь поверхность с двумя «дырами», изображенную на рис. 136. На этой поверхности оказывается возможным провести сразу две замкнутые кривые, A и B, которые не разделяют поверхности на части. Тор, напротив, при проведении двух таких кривых непременно разделится на части. С другой стороны, любые три замкнутые кривые разделяют нашу поверхность с двумя дырами.

Все это подсказывает мысль ввести понятие *рода* поверхности, понимая под родом поверхности наибольшее возможное число взаимно не пересекающихся простых замкнутых кривых, которые можно провести на поверхности, не разделяя ее на части. Род сферы равен 0, род тора равен 1, род поверхности, изображенной на рис. 136, равен 2. Такая же поверхность с p «дырами» имеет род p. Род есть топологический инвариант поверхности: он не изменяется при деформировании поверхности. Обратно, можно доказать (но мы не приводим здесь этого доказательства), что если две замкнутые поверхности имеют один и тот же род, то одну можно деформировать в другую; таким образом род $p = 0, 1, 2, \ldots$ замкнутой поверхности полностью характеризует топологический класс, к которому она принадлежит. (Здесь предполагается, что мы рассматриваем только обыкновенные «двусторонние» поверхности. В пункте 3 этого параграфа мы рассмотрим также и «односторонние» поверхности.) Например, только что рассмотренная поверхность с двумя дырами и сфера с двумя «рукоятками», изображенная на рис. 137, являются обе замкнутыми поверхностями рода 2, и мы видим, что каждую из этих поверхностей удается деформировать в другую. Так как поверхность с p дырами или ее эквивалент — сфера с p рукоятками — поверхности рода p, то любую из этих поверхностей можно взять в качестве «топологического представителя» всех замкнутых поверхностей рода p.

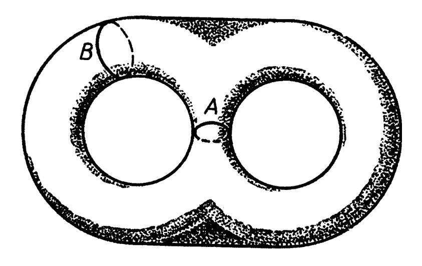


Рис. 136. Поверхность рода 2

***2. Эйлерова характеристика поверхности.** Предположим, что замкнутая поверхность S рода p разбита на некоторое число областей: такое подразделение получается, если мы отметим на S ряд «вершин» и соединим их затем между собой дугами кривых. Мы покажем, что в

таком случае

$$V - E + F = 2 - 2p, \tag{1}$$

где V — число вершин, E — число дуг и F — число областей. Число $2 - 2p$ называется *эйлеровой характеристикой* поверхности. Как мы уже видели, для случая сферы $V - E + F = 2$, что согласуется с формулой (1), так как сфера имеет род p, равный нулю.

Желая доказать общую формулу (1), вообразим, что S есть сфера с p рукоятками. Как мы отметили, любая поверхность рода p может быть непрерывной деформацией приведена к этому виду, и во время деформации ни $V - E + F$, ни $2 - 2p$ не изменяются. Непрерывную деформацию мы выберем таким образом, чтобы замкнутые кривые A_1, A_2, B_1, B_2, ..., по которым рукоятки соединяются со сферой, пришлись как раз на дуги данного подразделения. (Рис. 138 иллюстрирует описываемую дальше процедуру в случае $p = 2$.)

Прорежем теперь поверхность S по кривым A_2, B_2, ... и выпрямим рукоятки. У каждой рукоятки появится свободный край, ограниченный новой кривой A^*, B^*, ..., причем на появившемся крае будет столько же вершин и столько же дуг, сколько их было соответственно на A_2, B_2, ...

Число $V - E + F$ при прорезывании не изменится, так как новых областей не возникнет, а число вновь возникших вершин уравновешивается числом вновь возникших дуг. Затем деформируем поверхность дальше, сплющивая торчащие рукоятки (включая их в поверхность сферы). В итоге получается сфера с $2p$ отверстиями. Так как $V - E + F$, как нам известно, равно 2 для всякого разбиения полной сферы, то для нашей сферы с $2p$ отверстиями мы получаем $V - E + F = 2 - 2p$, и это равенство, очевидно, справедливо также и для первоначальной сферы с p рукоятками. Наше утверждение доказано.

Рис. 121 иллюстрирует применение формулы (1) к поверхности S, составленной из плоских многоугольников. Эту поверхность можно топологически деформировать в поверхность тора, так что ее род p равен 1, и потому $2 - 2p = 2 - 2 = 0$. Как и требуется по формуле (1), мы получаем

$$V - E + F = 16 - 32 + 16 = 0.$$

Упражнение. Произведите какое-нибудь разбиение на поверхности с двумя дырами, изображенной на рис. 137, и проверьте, что $V - E + F = -2$.

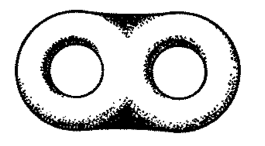
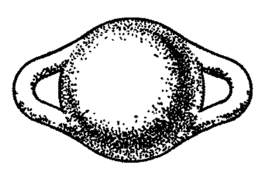
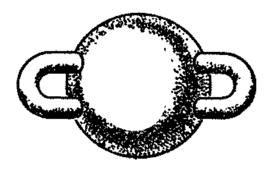




Рис. 137. Поверхности рода 2

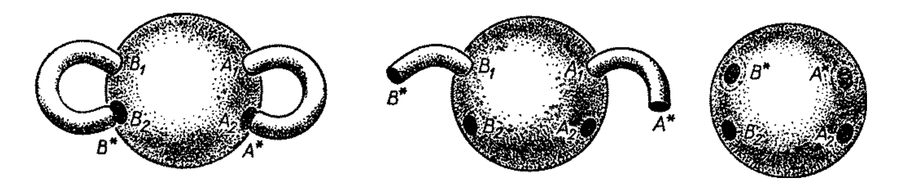


Рис. 138. К эйлеровой характеристике поверхностей

3. Односторонние поверхности. У каждой из обыкновенных поверхностей имеется по две стороны. Это относится и к замкнутым поверхностям вроде сферы или тора, и к поверхностям, имеющим границы, каковы, например, диск или тор, из которого удален кусок поверхности. Чтобы легко различать две стороны одной и той же поверхности, их можно было бы раскрасить разными красками. Если поверхность замкнутая, две краски нигде не встретятся. Если поверхность имеет граничные кривые, то разные краски встречаются по этим кривым. Предположим, что по таким поверхностям ползал бы клоп и что-нибудь мешало бы ему пересекать граничные кривые; тогда он оставался бы всегда на

Рис. 139. Лист Мёбиуса: *а*, *б*, *в* — перекручивание и склеивание ленты; *г* — ориентация «сторон»

одной стороне поверхности.

Мёбиусу принадлежит честь ошеломляющего открытия: существуют поверхности, у которых имеется только одна сторона. Простейшая из таких поверхностей есть так называемая лента (или лист) Мёбиуса. Чтобы ее построить, нужно взять лист бумаги, имеющий форму очень вытянутого прямоугольника, и склеить его концы после полуповорота, как показано на рис. 139 (*а*, *б*, *в*). Клоп, который будет ползти по этой поверхности, держась все время середины «ленты», вернувшись в исходную точку, окажется в перевернутом положении (рис. 139, *г*). Если кто-нибудь вздумает раскрасить «только одну» сторону поверхности мёбиусовой ленты, пусть лучше сразу погрузит ее всю в ведро с краской.

Другое замечательное свойство поверхности Мёбиуса заключается в том, что у нее только один край: вся граница состоит из одной замкнутой кривой. Обыкновенная двусторонняя поверхность, получающаяся при склеивании концов ленты без всякого поворота, явственно имеет две различные граничные кривые. Если эту последнюю поверхность разрезать по центральной линии, она распадется на две поверхности того же типа. Но если разрезать таким же образом по центральной линии ленту Мёбиуса (см. рис. 139), то мы увидим, что распадения на две части не будет. Тому, кто не упражнялся с лентой Мёбиуса, трудно предсказать это обстоятельство, столь противоречащее нашим интуитивным представлениям о том, что «должно» случиться. Но если поверхность, полученную после описанного выше разрезания ленты Мёбиуса, снова разрезать по ее центральной линии, то у нас в руках окажутся две не связанные, но переплетенные между собой ленты!

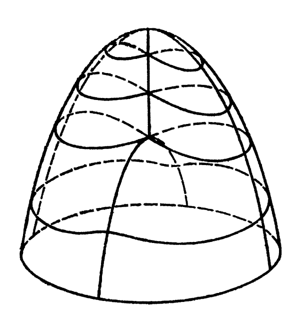


Рис. 140. Кросс-кэп

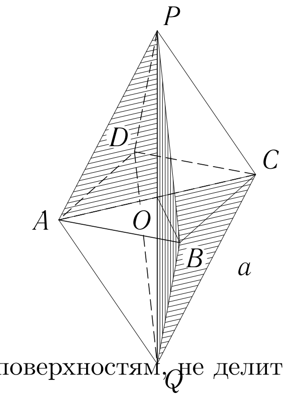
Очень интересно разрезать такие ленты по линиям, параллельным границе и отстоящим от нее на $\frac{1}{2}$, $\frac{1}{3}$ и т. д. ширины ленты. Поверхность Мёбиуса, без сомнения, заслуживает упоминания и в школьном курсе.

Граница поверхности Мёбиуса представляет собой простую «незаузленную» замкнутую кривую, и ее можно деформировать в окружность. Но придется допустить, что в процессе деформации поверхность будет сама себя пересекать. Получающаяся при этом самопересекающаяся односторонняя поверхность известна под названием «кросс-кэп» (рис. 140)[1]. Линию пересечения здесь следует считать дважды, один раз относя к одному из пересекающихся листов поверхности, другой раз — к другому. Кросс-кэп, как и всякую одностороннюю поверхность, нельзя непрерывно деформи-

[1] Cross-cap — «перекрещивающаяся шляпа» (англ.).

ровать в двустороннюю (топологическое свойство).

Любопытно, что ленту Мёбиуса можно, оказывается, так деформировать, что ее граница будет плоской ломаной (а именно, треугольником), причем лента останется несамопересекающейся. Такая модель, найденная д-ром Б. Туккерманом, показана на рис. 141, *а*; границей ленты служит треугольник ABC, ограничивающий половину диагонального квадратного сечения октаэдра (симметричного относительно этого сечения). Сама лента состоит при этом из шести граней октаэдра и четырех прямоугольных треугольников — четвертей вертикальных диагональных плоскостей октаэдра[1].



Другой любопытный пример односторонней поверхности — так называемая «бутылка Клейна». Это — замкнутая поверхность, но она, в противоположность известным нам замкнутым поверхностям, не делит

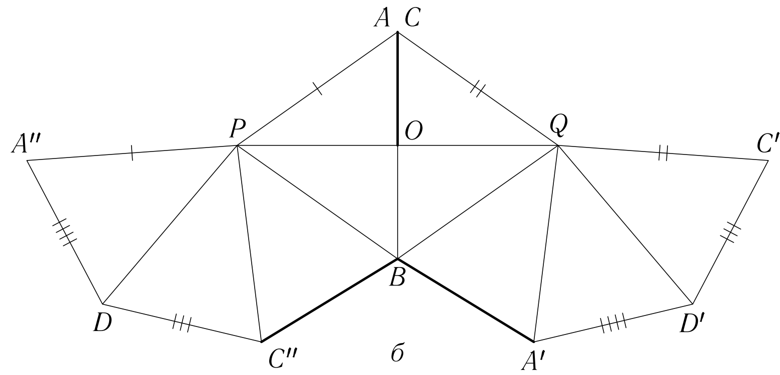


Рис. 141. Лента Мёбиуса с прямолинейным краем (*а*) и ее развертка (*б*)

пространство на «внутреннюю» и «внешнюю» части. Топологически она эквивалентна паре кросс-кэпов со склеенными между собой граничными кривыми.

Можно доказать, что всякая замкнутая *односторонняя* поверхность рода $p = 1, 2, \ldots$ топологически эквивалентна сфере, из которой вынуты p дисков и заменены кросс-кэпами. Отсюда легко выводится, что

[1] Из поверхности октаэдра вырезаются грани ABP и BCQ. К оставшимся шести граням приклеиваются четыре треугольника OAP, OBP, OCQ и OBQ. На рис. 141, *б* приведена развертка описанной поверхности. По линии, соединяющей точку O с точкой, помеченной двумя буквами A и C, надо сделать разрез, а потом склеить соответствующие отрезки края развертки. Жирными отрезками обозначен край ленты (периметр треугольннка ABC). — *Прим. ред.*

эйлерова характеристика $V - E + F$ такой поверхности связана с родом p соотношением

$$V - E + F = 2 - p.$$

Доказательство этого предложения такое же, как и для двусторонних поверхностей. Прежде всего убедимся, что Эйлерова характеристика кросс-кэпа или ленты Мёбиуса равна 0. Для этого заметим, что, перерезая поперек ленту Мёбиуса, предварительно подразделенную на области, мы получим прямоугольник, у которого будут две лишние вершины и одна лишняя дуга, число же областей останется то же самое, что и для ленты Мёбиуса. Мы видели на стр. 260, что для прямоугольника $V - E + F = 1$. Следовательно, для ленты Мёбиуса $V - E + F = 0$. Предлагаем читателю в качестве упражнения восстановить это доказательство во всех подробностях.

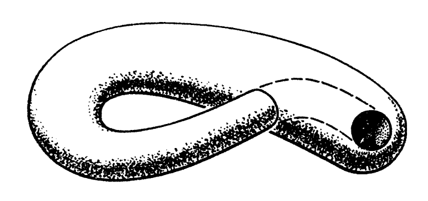


Рис. 142. Бутылка Клейна

Изучение топологической структуры поверхностей, подобных тем, которые только что были описаны, проводится более удобно, если воспользоваться плоскими многоугольниками с попарно идентифицированными сторонами (см. гл. IV, Приложение, пункт 3). Так, на схемах рис. 143 стрелки показывают, какие из параллельных сторон и в каком направлении должны быть идентифицированы: если возможно, то физически, если невозможно, то хотя бы мысленно, абстрактно.

Метод идентификации можно применить и для определения трехмерных замкнутых многообразий, аналогичных двумерным замкнутым поверхностям. Например, отождествляя соответствующие точки взаимно противоположных граней куба (рис. 144), мы получаем замкнутое трехмерное многообразие, называемое трехмерным тором. Такое многообразие топологически эквивалентно пространственной области, заключенной между двумя концентрическими поверхностями тора (одна внутри другой), с идентификацией соответствующих точек (рис. 145). Действительно, это последнее многообразие получается из куба, если привести в «физическое»

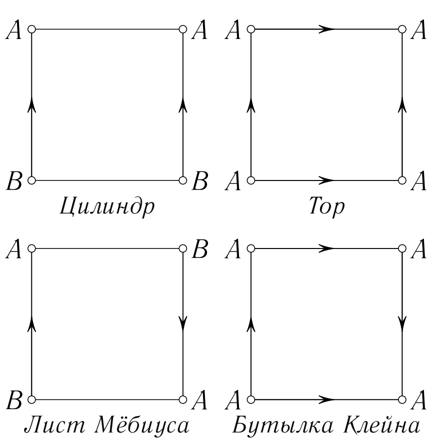


Рис. 143. Замкнутые поверхности, определенные посредством идентификации сторон квадрата

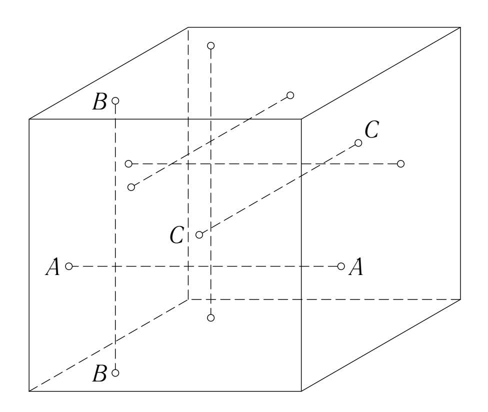


Рис. 144. Определение трехмерного тора посредством идентификации граней куба

совпадение две пары «мысленно отождествленных» взаимно противоположных граней.

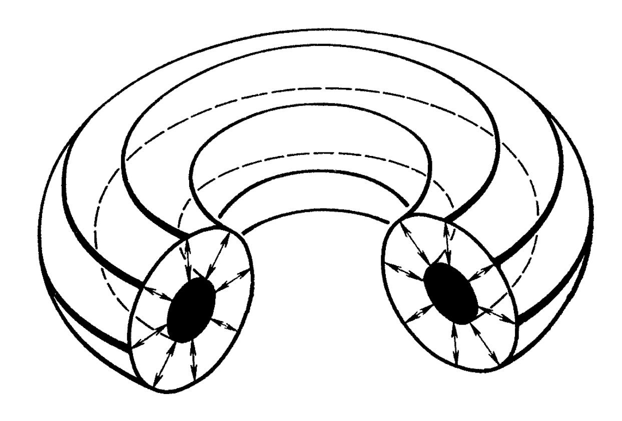


Рис. 145. Другое представление трехмерного тора (разрезы показывают идентификацию)

ПРИЛОЖЕНИЕ

***1. Проблема пяти красок.** Доказательство того, что всякая карта на сфере может быть правильно раскрашена с помощью не более чем пяти различных красок, основывается на формуле Эйлера. (В соответствии с тем, что было сказано на стр. 265, карта считается раскрашенной правильно, если никакие две области, соприкасающиеся по целой дуге, не окрашены одинаково.) Мы ограничимся рассмотрением таких карт, на которых все области являются простыми замкнутыми многоугольниками, ограниченными круговыми дугами. Не уменьшая общности, можно допустить, что в каждой вершине сходится ровно по три дуги; карту, удовлетворяющую такому условию, будем называть регулярной. Заменим каждую вершину, в которой встречается более трех дуг, маленьким кружком и присоединим внутренность этого кружка к одной из прилегающих областей: тогда получится новая карта, в которой «кратные» вершины заменены обыкновенными. Новая карта будет содержать столько же областей, сколько и старая, и она будет регулярной. Если ее удастся правильно раскрасить, то, сжимая потом кружок и сводя его в точку, мы получим требуемую раскраску первоначальной карты. Таким образом, достаточно убедиться, что каждую регулярную карту на сфере можно правильно раскрасить пятью красками.

Прежде всего докажем, что всякая регулярная карта содержит по крайней мере один многоугольник с числом сторон меньшим шести. Обозначим через F_n число n-угольных областей нашей регулярной карты; тогда, обозначая через F число всех областей, получим равенство

$$F = F_2 + F_3 + F_4 + \ldots \tag{1}$$

У каждой дуги два конца; в каждой вершине сходится три дуги. Поэтому если E обозначает число дуг, а V — число вершин на нашей карте, то

$$2E = 3V. \tag{2}$$

Далее, так как область, ограниченная n дугами, имеет n вершин и каждая вершина принадлежит трем областям, то

$$2E = 3V = 2F_2 + 3F_3 + 4F_4 + \ldots \tag{3}$$

По формуле Эйлера мы имеем

$$V - E + F = 2, \quad \text{или} \quad 6V - 6E + 6F = 12.$$

Из соотношения (2) следует, что $6V = 4E$, так что $6F - 2E = 12$. Тогда соотношения (1) и (3) нам дают

$$6(F_2 + F_3 + F_4 + \ldots) - (2F_2 + 3F_3 + 4F_4 + \ldots) = 12,$$

или

$$(6-2)F_2 + (6-3)F_3 + (6-4)F_4 + (6-5)F_5 + \\ + (6-6)F_6 + (6-7)F_7 + \ldots = 12.$$

Так как хотя бы один из членов суммы слева должен быть положительным, то отсюда ясно, что четыре числа F_2, F_3, F_4 и F_5 не могут одновременно равняться нулю. Но это и есть то, что нам нужно было доказать.

Перейдем теперь к теореме о пяти красках. Пусть M — какая угодно регулярная карта на сфере, n — число всех ее областей. Мы знаем, что имеется хоть одна область с числом сторон меньшим шести.

Случай 1. M содержит область A с 2, 3 или 4 сторонами (рис. 146). В этом случае удалим дугу, отделяющую A от одной из прилежащих областей. (Здесь необходимо следующее примечание. Если у области A четыре стороны, то может случиться, что одна из прилежащих областей, если ее обойти вокруг, окажется также прилежащей к A с противоположной стороны. В этом случае, на основании теоремы Жордана, две области, прилегающие к A с двух остальных сторон, будут различными, и мы сможем удалить одну из этих двух сторон.) Вновь полученная карта M' будет также регулярной, но содержащей уже только $n-1$ областей. *Если карту M' можно правильно раскрасить пятью красками, то можно раскрасить и карту M.* В самом деле, к области A прилегает самое большее четыре области карты M, и потому для A всегда найдется пятая краска.

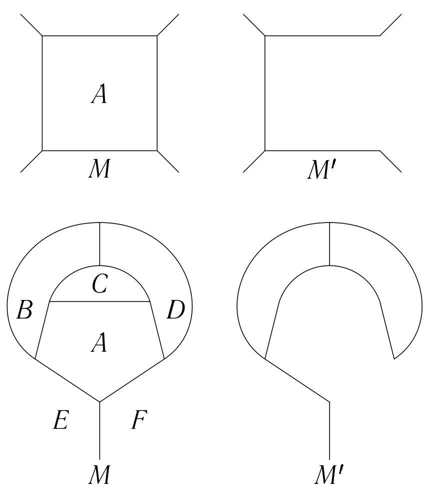


Рис. 146–147. К доказательству теоремы о пяти красках

Случай 2. M содержит область A с пятью сторонами. Рассматривая пять областей, прилегающих к A, обозначим их через B, C, D, E и F. Можно всегда среди этих пяти областей найти две, которые не прилегают друг к другу: действительно, если, например, B и D прилегают одна к другой, то отсюда вытекает, что C не прилегает ни к E, ни к F, так как в противном случае всякий путь, идущий из C в E или F, должен был бы пройти по крайней мере через одну из областей A, B или D (рис. 147). (Ясно, что это утверждение существенно зависит от теоремы Жордана, справедливой для плоскости и для сферы. Для тора, напротив, все это рассуждение отпадает.) Итак, можно допустить, что, например, C и F не прилегают друг к другу. Удалим те две стороны A, которые отделяют A от C и F, и тогда получим новую карту M' с $n-2$ областями, также регулярную. *Если новую карту M' можно правильно раскрасить пятью*

красками, то можно раскрасить и первоначальную карту M. В самом деле, после восстановления удаленных сторон область A будет прилегать к пяти областям, окрашенным не более чем четырьмя красками (так как C и F окрашены одинаково), и потому для A всегда найдется пятая краска.

Таким образом, во всех случаях всякий регулярной карте M с n областями можно сопоставить такую, также регулярную, карту M' с $n-1$ или $n-2$ областями, что если M' можно раскрасить пятью красками, то M — также. Это рассуждение можно повторить для карты M' и т. д. В результате мы получим последовательность карт, в которой первым членом явится M:

$$M,\ M',\ M'',\ \ldots,$$

обладающую тем свойством, что каждая карта этой последовательности может быть раскрашена пятью красками, если может быть раскрашена следующая за ней. Но так как число областей в картах этой последовательности неизменно убывает, то рано или поздно в ней найдется карта с пятью областями (или меньшим числом). Такую карту всегда можно раскрасить не более чем пятью красками. Возвращаясь по последовательности карт, шаг за шагом, обратно, заключим отсюда, что и сама карта M может быть раскрашена пятью красками. Этим доказательство и заканчивается.

Остается заметить, что это доказательство имеет «конструктивный» характер: оно дает практически осуществимый, хотя, может быть, и утомительный, метод нахождения требуемой раскраски данной карты M, составленной из n областей, посредством конечного числа шагов.

2. Теорема Жордана для случая многоугольников. Теорема Жордана утверждает, что всякая простая замкнутая кривая C разделяет точки плоскости, не принадлежащие C, на такие две области (не имеющие общих точек), по отношению к которым сама кривая C является общей границей. Мы докажем здесь эту теорему для частного случая, когда C есть замкнутый многоугольник P.

Мы покажем, что точки плоскости (кроме точек, находящихся на самом многоугольном контуре P) разбиваются на два класса A и B, обладающие следующими свойствами: 1) две точки одного и того же класса могут быть соединены ломаной линией, не имеющей общих точек с P; 2) если две точки принадлежат разным классам, то любая ломаная линия, их соединяющая, имеет общие точки с P. Один из названных классов образует «внутренность» многоугольника, другой состоит из точек, находящихся «вне» многоугольника.

Приступая к доказательству, выберем какое-то фиксированное направление в нашей плоскости, не параллельное ни одной из сторон P. Так как P имеет конечное число сторон, то это всегда возможно. Затем

определим классы A и B следующим образом.

Точка p принадлежит классу A, если луч, проведенный через нее в фиксированном направлении, пересекает P в *четном* числе точек (0, 2, 4, 6 и т. д.). Точка p принадлежит классу B, если луч, проведенный из p в фиксированном направлении, пересекает P в *нечетном* числе точек (1, 3, 5 и т. д.).

К этому нужно добавить, что если рассматриваемый луч проходит через какую-нибудь вершину, то эта вершина идет в счет как точка пересечения луча с P или не идет, смотря по тому, расположены ли прилежащие стороны многоугольника P по разные стороны луча или по одну и ту же его сторону.

Условимся говорить, что две точки p и q имеют одну и ту же «четность», если они принадлежат одному и тому же из двух классов A и B.

Заметим прежде всего, что все точки любого отрезка прямой, не пересекающегося с P, имеют одну и ту же четность. Действительно, четность точки p, движущейся по такому отрезку, может измениться не иначе, как при пересечении соответствующего луча с одной из вершин P; но, принимая во внимание наше соглашение о счете точек пересечения, легко убедиться, что в каждом из двух возможных случаев четность все же не меняется. Из сказанного следует, что *если некоторая точка p_1 области A соединена ломаной линией с некоторой точкой p_2 области B, то эта линия непременно пересекает P.* Иначе четность всех точек ломаной линии, в частности, точек p_1 и p_2, была бы одинаковой. Дальше, покажем, что *две точки одного и того же из двух классов A и B могут быть соединены ломаной линией, не пересекающейся с P.* Обозначим две данные точки через p и q. Если прямолинейный отрезок pq, соединяющий p и q, не пересекается с P, то доказывать больше нечего. В противном случае пусть p' — первая, а q' — последняя точка пересечения отрезка pq с многоугольником P (рис. 149). Построим ломаную линию, начинающуюся от точки p прямолинейным отрезком, расположенным по направлению pq, но заканчивающуюся непосредственно перед точкой p': отсюда ломаная пойдет вдоль P (безразлично, в каком из двух возможных направлений) и будет так идти, пока не придет снова на

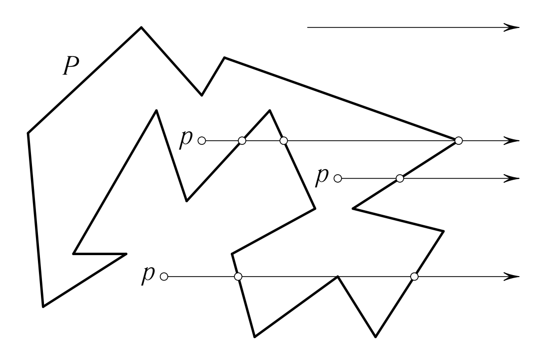


Рис. 148. Счет пересечений

прямую pq — около точки q'. Весь вопрос в том, произойдет ли пересечение с прямой pq на отрезке $p'q'$ или на отрезке $q'q$: мы сейчас убедимся, что справедливо именно последнее, и тогда будем иметь возможность закончить ломаную, соединяя последнюю из полученных точек с точкой q прямолинейным отрезком, снова лежащим на отрезке pq. Если две точки r и s расположены очень близко одна от другой, но по разные стороны одной из сторон многоугольника P, то они имеют различную четность, так как выходящие из них (в фиксированном направлении) лучи будут таковы, что на одном из них будет на одну точку больше точек пересечения с P, чем на другом. Отсюда ясно, что четность меняется, когда, двигаясь по pq, мы проходим через точку q'. Значит, ломаный «путь», намеченный на чертеже пунктиром, вернется на pq между q' и q, так как p и q (следовательно, все точки на рассматриваемом «пути») имеют одну и ту же четность.

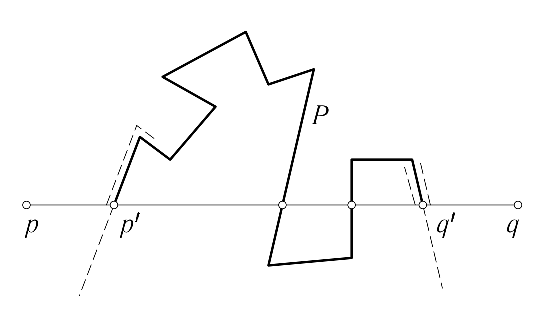


Рис. 149. К доказательству теоремы Жордана

Таким образом, теорема Жордана для случая многоугольника доказана. «Внешними» по отношению к многоугольнику P будут те точки, которые принадлежат классу A: действительно, двигаясь по какому-нибудь лучу в фиксированном направлении достаточно далеко, мы, несомненно, придем к точке, за которой пересечений с P уже не будет, и все такие точки будут принадлежать классу A, так что их четность будет 0. Тогда уже придется заключить, что точками «внутренними» будут точки класса B. Каким бы запутанным ни был замкнутый многоугольник P, всегда очень легко узнать, расположена ли данная точка p внутри или вне его: достаточно из p провести луч и посчитать число его точек пересечения с P. Если это число нечетное, значит, p «сидит» внутри и не сможет выбраться наружу, не пересекая P. Но если это число четное, то точка p — вне многоугольника P (попробуйте проверить это на рис. 128).

Вот идея другого доказательства жордановой теоремы для случая многоугольников. Определим *порядок* точки p_0 относительно замкнутой кривой C (не проходящей через p_0) как число полных поворотов[1], совершаемых стрел-

[1] Это число нужно, конечно, брать в алгебраическом смысле, т. е. с учетом направления вращения.

кой (вектором), проведенным от p к p_0, когда точка p делает один обход по кривой C.

Затем пусть A есть совокупность точек p_0 (не на P) *четного* порядка относительно P, а B есть совокупность точек p_0 (не на P) *нечетного* порядка относительно P. В таком случае A и B, определенные указанным способом, представляют собой соответственно области «внешнюю» и «внутреннюю» относительно P. Читатель может в качестве упражнения воспроизвести все детали доказательства.

***3. Основная теорема алгебры.** Основная теорема алгебры утверждает, что если функция $f(z)$ имеет вид

$$f(z) = z^n + a_{n-1}z^{n-1} + a_{n-2}z^{n-2} + \ldots + a_1 z + a_0, \tag{1}$$

где $n \geqslant 1$ и a_{n-1}, a_{n-2}, ..., a_1, a_0 — какие угодно комплексные числа, то существует такое комплексное число α, что $f(\alpha) = 0$. Другими словами, *в поле комплексных чисел всякое алгебраическое уравнение имеет корень*. (Основываясь на этой теореме, мы на стр. 122 сделали дальнейшее заключение: полином $f(z)$ может быть разложен на n линейных множителей

$$f(z) = (z - \alpha_1)(z - \alpha_2)\ldots(z - \alpha_n),$$

где α_1, α_2, ..., α_n — нули $f(z)$.) Замечательно, что эту теорему можно доказать, исходя из соображений топологического характера, как и теорему Брауэра о неподвижной точке.

Пусть читатель вспомнит, что комплексное число есть символ вида $x + yi$, где x и y — действительные числа, а символ i обладает свойством $i^2 = -1$. Комплексное число $x + yi$ изображается точкой (x, y) в плоскости прямоугольных координат. Если мы введем в этой же плоскости полярные координаты, принимая начало координат за полюс, а положительное направление оси x за полярную ось, то можно будет написать

$$z = x + yi = r(\cos\theta + i\sin\theta),$$

где $r = \sqrt{x^2 + y^2}$. Из формулы Муавра следует, что $z^n = r^n(\cos n\theta + i\sin n\theta)$ (см. стр. 118). Отсюда ясно, что если комплексное число z описывает круг радиуса r с центром в начале координат, то z^n опишет ровно n раз круг с радиусом r^n. Напомним еще, что модуль z (обозначаемый через $|z|$) представляет собой расстояние z от 0 и что если $z' = x' + iy'$, то $|z - z'|$ есть расстояние между z и z'. После этих напоминаний можно перейти к доказательству теоремы.

Допустим, что полином (1) *не* имеет корней, так что при любом комплексном z

$$f(z) \neq 0.$$

При этом допущении, если z описывает некоторую замкнутую кривую в плоскости x, y, то $f(z)$ опишет некоторую замкнутую кривую Γ, не

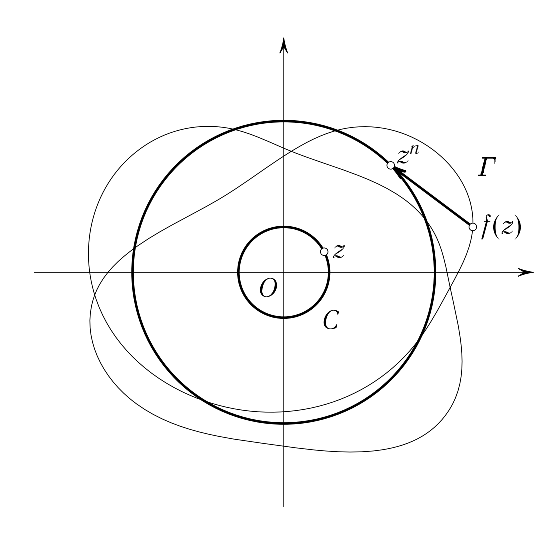


Рис. 150. Доказательство основной теоремы алгебры

проходящую через начало координат (рис. 150). Можно определить *порядок* точки O для функции $f(z)$ относительно замкнутой кривой C как *число полных поворотов, совершаемых вектором, идущим от O к точке $f(z)$ на кривой Γ, когда z делает полный обход по кривой C*. Возьмем в качестве кривой C окружность с центром O и радиусом t и обозначим через $\varphi(t)$ порядок точки O для функции $f(z)$ относительно окружности с центром O и радиусом t. Очевидно, $\varphi(0) = 0$, так как круг радиуса 0 сводится к одной точке и кривая Γ также сводится к одной точке $f(0) \neq 0$. Если мы докажем, что при достаточно больших значениях t функция $\varphi(t)$ равна n, то в этом уже будет заключаться противоречие, так как, с одной стороны, порядок $\varphi(t)$ должен быть непрерывной функцией t (поскольку $f(z)$ есть непрерывная функция z), а с другой стороны, функция $\varphi(t)$ может принимать только целые значения и потому никак не может перейти от значения 0 к значению n непрерывно.

Нам остается доказать, что при достаточно больших значениях t

$$\varphi(t) = n.$$

Для этого заметим, что если радиус круга t удовлетворяет неравенствам

$$t > 1 \quad \text{и} \quad t > |a_0| + |a_1| + \ldots + |a_{n-1}|,$$

то

$$\begin{aligned}|f(z) - z^n| &= |a_{n-1}z^{n-1} + a_{n-2}z^{n-2} + \ldots + a_0| \leqslant \\ &\leqslant |a_{n-1}| \cdot |z|^{n-1} + |a_{n-2}| \cdot |z|^{n-2} + \ldots + |a_0| = \\ &= t^{n-1}\left[|a_{n-1}| + \frac{|a_{n-2}|}{t} + \ldots + \frac{|a_0|}{t^{n-1}}\right] \leqslant \\ &\leqslant t^{n-1}\big[|a_{n-1}| + |a_{n-2}| + \ldots + |a_0|\big] < t^n = |z^n|.\end{aligned}$$

Так как выражение слева есть не что иное, как расстояние между точками z^n и $f(z)$, а выражение в самой правой части неравенства — расстояние точки z^n от начала координат, то мы видим отсюда, что отрезок, соединяющий точки z^n и $f(z)$, не пройдет через начало координат, если только z будет находиться на круге радиуса t с центром в начале. В таком случае имеется возможность *деформировать* кривую, описываемую точкой $f(t)$, в кривую, описываемую точкой z^n, без прохода через начало, смещая непрерывным движением каждую точку $f(z)$ к соответствующей точке z^n по прямоугольному отрезку. При этом порядок начала может принимать только целые значения и вместе с тем во время деформации может меняться не иначе, как непрерывно; значит, для обеих функций $f(z)$ и z^n он одинаков, и так как для z^n он равен n, то имеет то же самое значение и для $f(z)$. Доказательство закончено.

ГЛАВА VI

Функции и пределы

Введение

Важнейшие разделы современной математики сосредоточиваются вокруг понятий функции и предела. В этой главе мы займемся систематическим анализом этих понятий.

Такие выражения, как, например,

$$x^2 + 2x - 3,$$

не имеют определенного числового значения, пока не указано значение x. Говорят, что значение этого выражения есть *функция* значения x, и пишут

$$x^2 + 2x - 3 = f(x).$$

Например, если $x = 2$, то $2^2 + 2 \cdot 2 - 3 = 5$, так что $f(2) = 5$. Таким же образом непосредственной подстановкой можно найти значение функции $f(x)$ при любом целом, дробном, иррациональном и даже комплексном значении x.

Количество простых чисел, меньших чем n, есть функция $\pi(n)$ целого числа n. Когда задано значение числа n, то значение функции $\pi(n)$ определено, несмотря на то что неизвестно никакого алгебраического выражения для его подсчета. Площадь треугольника есть функция длин трех его сторон; она меняется вместе с ними и делается фиксированной, если зафиксированы длины сторон. Если плоскость подвергается проективному или топологическому преобразованию, то координаты точки после преобразования зависят от первоначальных координат точки, т. е. являются их функциями. Понятие «функция» выступает каждый раз, как только величины связаны каким-нибудь определенным физическим соотношением. Объем газа, заключенного в цилиндр, есть функция температуры и давления, оказываемого на поршень. Замечено, что давление атмосферы на воздушный шар есть функция высоты шара над уровнем моря. Целая область периодических явлений — движение приливов, колебание натянутой струны, распространение световых волн, испускаемых накаленной проволокой, — «регулируется» простыми тригонометрическими функциями $\sin x$ и $\cos x$.

Для самого Лейбница (1646–1716), который впервые ввел термин «функция», и для математиков XVIII в. идея функциональной зависимости более или менее идентифицировалась с существованием простой математической формулы, точно выражающей эту зависимость. Такая концепция оказалась слишком узкой по отношению к требованиям, предъявленным математической физикой, и понятие «функция» вместе с упомянутым выше понятием «предел» впоследствии длительно подвергалось обобщениям и шлифовке.

В этой главе мы дадим краткий очерк того, как протекал этот процесс.

§ 1. Независимое переменное и функция

1. Определения и примеры. Нередко приходится иметь дело с математическими объектами, которые мы выбираем свободно, по нашему собственному выбору, из некоторой совокупности (множества) S. Избираемый объект в таких случаях носит название *переменного* (или *переменной*), а совокупность S — *области его* (*ее*) *изменения*. Переменные принято обозначать последними буквами алфавита. Например, если буквой S обозначено множество всех целых чисел, то переменное X из области S обозначает некоторое произвольное целое число. Говорят, что «переменное X пробегает множество S», подразумевая под этим, что переменное X мы можем отождествить с любым элементом множества S. Пользоваться понятием переменного удобно, если мы хотим высказать утверждение относительно элементов, которые можно произвольно выбирать из целого множества. Например, если S обозначает, как было указано, множество целых чисел, а X и Y — переменные из области S, то формула

$$X + Y = Y + X$$

представляет удобное символическое выражение того обстоятельства, что сумма любых двух целых чисел не зависит от порядка слагаемых. Частный случай этого выражен равенством

$$2 + 3 = 3 + 2,$$

в котором фигурируют постоянные числа; но для того чтобы выразить общий закон, справедливый для всех пар чисел, нужно применить символы, имеющие значение переменных.

Нет никакой необходимости в том, чтобы область S изменения переменного X была множеством чисел. Например, S может быть множеством всех кругов на плоскости; тогда переменное X будет обозначать любой индивидуальный круг. Или S может быть множеством всех замкнутых многоугольников плоскости, и тогда X — любой индивидуальный многоугольник. Не является также необходимым, чтобы область изменения переменного содержала бесконечное число элементов. Например, X

может обозначать любого отдельного человека из населения S данного города в определенный момент времени. Или же X может обозначать любой из возможных остатков при делении целого числа на 5; в этом последнем случае область S состоит из пяти чисел: 0, 1, 2, 3, 4.

Наиболее важным оказывается случай числового переменного; в этом случае употребляется обычно маленькая буква x — это тот случай, когда областью изменения S является некоторый интервал (промежуток) $a \leqslant \leqslant x \leqslant b$ действительной числовой оси. В этом случае говорят, что x есть *непрерывное* (или *действительное*) *переменное* в рассматриваемом интервале. Область изменения непрерывного переменного может простираться и до бесконечности. Так, например, S может быть множеством всех положительных действительных чисел $x > 0$ или даже множеством *всех* действительных чисел без всякого исключения. Аналогичным образом мы можем рассматривать переменное X, значениями которого являются точки плоскости или некоторой данной области плоскости, подобной внутренности прямоугольника или круга. Так как каждая точка плоскости определяется своими двумя координатами (x, y), взятыми относительно некоторой фиксированной пары осей, то в этом случае часто говорят, что имеют дело с *парой действительных* (*непрерывных*) *переменных* x и y.

Может случиться так, что каждому значению переменного X сопоставляется некоторое определенное значение другого переменного U. Тогда переменное U называется *функцией* переменного X. Способ, посредством которого U связано с X, выражается символом вроде $U = F(X)$ (читается «равно F от X»). Если X пробегает множество S, то переменное U пробегает некоторое другое множество, скажем, T. Например, если S есть множество треугольников X на плоскости, то под функцией $U = F(X)$ можно подразумевать длину периметра рассматриваемого треугольника X; T будет, следовательно, множеством всех положительных чисел. Отметим, что два различных треугольника X_1 и X_2 свободно могут иметь равные по длине периметры, так что равенство $F(X_1) = F(X_2)$ возможно и в том случае, если $X_1 \neq X_2$. Проективное преобразование одной плоскости S в некоторую другую T ставит в соответствие каждой точке X плоскости S единственную точку U плоскости T согласно определенному правилу, которое можно выразить функциональным символом $U = F(X)$. В этом примере, в противоположность предыдущему, мы имеем всегда неравенство $F(X_1) \neq F(X_2)$, если только $X_1 \neq X_2$, и мы говорим в связи с этим, что отображение плоскости S на плоскость T — *взаимно однозначное* (см. стр. 204).

Функции непрерывного переменного часто определяются с помощью алгебраических выражений. Примерами могут служить следующие

функции:

$$u = x^2, \qquad u = \frac{1}{x}, \qquad u = \frac{1}{1+x^2}.$$

В первом и в последнем из этих выражений x может пробегать множество всех действительных чисел, в то время как во втором примере x может пробегать множество всех действительных чисел за исключением 0 (значение 0 исключается, так как символ $\frac{1}{0}$ не есть число).

Число $B(n)$ простых множителей числа n есть функция n, причем n пробегает множество натуральных чисел. Вообще, любую последовательность чисел a_1, a_2, a_3, ... можно рассматривать как множество значений некоторой функции $u = F(n)$, причем областью изменения независимого переменного при этом является множество натуральных чисел. Только ради сокращения записи принято обозначать n-й член последовательности символом a_n, вместо того чтобы употреблять более отчетливое функциональное обозначение $F(n)$. Следующие выражения, о которых говорилось в главе I:

$$S_1(n) = 1 + 2 + \ldots + n = \frac{n(n+1)}{2},$$

$$S_2(n) = 1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6},$$

$$S_3(n) = 1^3 + 2^3 + \ldots + n^3 = \frac{n^2(n+1)^2}{4},$$

являются функциями целого переменного n.

Пусть дано соотношение $U = F(X)$; принято переменное X называть *независимым переменным*, а переменное U — *зависимым*, поскольку его значения зависят от выбора значения X.

Может случиться, что всем значениям переменного X соответствует одно и то же значение переменного U, т. е. что множество T состоит из одного-единственного элемента. Мы тогда встречаемся с частным случаем, при котором переменное U в сущности не меняется, т. е. U есть постоянное (*постоянная* или *константа*). Мы включим этот случай в общее понятие функции, несмотря на то что начинающему это может показаться странным, так как он склонен полагать, что основное в самой идее функции лежит как раз в изменении переменного U (при изменении переменного X). Но беды не произойдет — а на самом деле это окажется весьма полезным, — если мы постоянное будем рассматривать как частный случай переменного, «область изменения» которого состоит из одного-единственного элемента.

Понятие функциональной зависимости имеет исключительное значение не только в самой «чистой» математике, но также и в практических ее приложениях. Физические законы являются не чем иным, как выражением способа, посредством которого некоторые величины зависят

от других, способных изменяться так или иначе. Так, например, высота звука, производимого колеблющейся струной, зависит от ее длины, от ее веса и от степени ее натяжения; давление атмосферы зависит от высоты; энергия пули зависит от ее массы и скорости. Задача физики состоит в точном или приближенном определении природы всех подобного рода зависимостей.

С помощью понятия функции можно дать точную в математическом смысле характеристику движения. Если представим себе, что движущаяся частица сосредоточена в некоторой точке пространства с прямоугольными координатами x, y, z, и если переменное t измеряет время, то движение частицы полностью определено заданием координат x, y, z как функций времени:

$$x = f(t), \qquad y = g(t), \qquad z = h(t).$$

Примером этого может служить свободное падение частицы по вертикали под действием одной лишь силы тяжести: мы имеем в этом случае соотношения

$$x = 0, \qquad y = 0, \qquad z = -\frac{1}{2}\,gt^2,$$

где g обозначает ускорение силы тяжести. Если частица равномерно вращается по единичной окружности в плоскости x, y, то движение ее характеризуется функциями

$$x = \cos\omega t, \qquad y = \sin\omega t,$$

где ω — постоянное число (так называемая угловая скорость вращения).

Под математической функцией следует понимать просто закон, управляющий взаимными зависимостями переменных величин — и не более того. Понятие функции не подразумевает существования чего-либо близкого к «причине и следствию» в отношениях между независимой и зависимой переменными. Хотя в обыденной речи термин «функциональная зависимость» сплошь и рядом употребляется именно в этом последнем смысле, мы будем избегать такого рода философских интерпретаций. Так, например, закон Бойля, относящийся к газу, заключенному в некоторую замкнутую оболочку при постоянной температуре, утверждает, что произведение давления газа p на его объем v есть величина постоянная, равная c (последнее значение, в свою очередь, зависит от температуры):

$$pv = c.$$

Это соотношение можно решить как относительно p, так и относительно v:

$$p = \frac{c}{v} \quad \text{или} \quad v = \frac{c}{p};$$

при этом не следует подразумевать ни того, что перемена объема есть «причина» изменения давления, ни того, что изменение давления есть

«причина» изменения объема. Для математика существенна лишь форма *соответствия* (*связи*) между двумя переменными величинами, которые он рассматривает.

Следует заметить, что подход к понятию функции несколько отличается у математиков и у физиков. Математики обычно подчеркивают *закон соответствия*, математическую операцию, которую нужно применить к значению независимого переменного x, чтобы получить значение зависимого переменного u. В этом смысле $f(\ )$ есть символ *математической операции*; значение $u = f(x)$ есть результат применения операции $f(\ )$ к числу x. С другой стороны, физик часто более заинтересован в *самой величине* u как таковой, чем в какой-то математической процедуре, с помощью которой значение u может быть получено из значения x. Так, например, сопротивление u воздуха движению предмета зависит от скорости v движения и может быть найдено экспериментальным путем, независимо от того, известна ли явная математическая формула для вычисления u. Физика прежде всего интересует фактическое сопротивление, а не специальная математическая формула $f(v)$, если только эта формула не помогает при анализе поведения величины u. Таково обычно отношение тех, кто *применяет* математику к физике или инженерному делу. В некоторых высших разделах математического анализа, чтобы избежать путаницы, иногда бывает существенно различать совершенно отчетливо, будет ли под символом $u = f(x)$ подразумеваться операция $f(\ )$, применяемая к x для получения u, или же сама величина u, которая, в свою очередь, может рассматриваться как зависимая, и совсем другим образом, от некоторой другой переменной z. Например, площадь круга задается функцией $u = f(x) = \pi x^2$, где x — радиус круга, но можно также написать: $u = g(z) = \frac{z^2}{4\pi}$, понимая под z длину окружности.

Пожалуй, наиболее простым типом математической функции одного независимого переменного являются *многочлены* (*полиномы*), имеющие вид

$$u = f(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n,$$

с постоянными «коэффициентами» a_0, a_1, $\ldots$, a_n. За ними следуют *рациональные функции*, такие как

$$u = \frac{1}{x}, \qquad u = \frac{1}{1+x^2}, \qquad u = \frac{2x+1}{x^4+3x^2+5},$$

которые являются отношениями многочленов, и затем *тригонометрические функции* $\cos x$, $\sin x$ и $\operatorname{tg} x = \frac{\sin x}{\cos x}$, которые определяются лучше всего с помощью единичного круга в плоскости ξ, η: $\xi^2 + \eta^2 = 1$. Если точка $P(\xi, \eta)$ движется по этой окружности и если x есть направленный угол, на который нужно повернуть положительную ось x, чтобы она совпала с радиусом OP, то $\cos x$ и $\sin x$ являются координатами точки P: $\cos x = \xi$, $\sin x = \eta$.

2. Радианная мера углов. Во всех практических применениях углы измеряются с помощью единиц, полученных от деления прямого угла на некоторое равное число частей. Если это число равно 90, то единицей измерения является обычный «градус». Деление на 100 частей подходило бы близко к нашей десятичной системе, но принцип измерения при этом оставался бы прежним. В теоретических же применениях выгоднее использовать по существу совершенно другой метод определения величины угла, а именно, так называемое радианное измерение. Многие важные формулы, содержащие тригонометрические функции углов, имеют в этой системе измерения более простой вид, чем при измерении углов в градусах.

Для того чтобы найти радианную меру некоторого угла, опишем из вершины этого угла как из центра круг радиуса 1.

Длину дуги s той части нашей окружности, которая расположена между сторонами угла, назовем *радианной мерой* угла. Так как длина всей окружности единичного радиуса равна 2π, то «полный» угол в 360° имеет радианную меру 2π. Отсюда следует, что если через x обозначить радианную меру угла, а через y его величину в градусах, то x и y связаны соотношением $\frac{y}{360} = \frac{x}{2\pi}$, или

$$\pi y = 180x.$$

Так, например, угол в 90° $(y = 90)$ имеет радианной мерой $x = \frac{90\pi}{180} = \frac{\pi}{2}$, и т. д. С другой стороны, угол в 1 радиан (угол, радианной мерой которого является $x = 1$) есть центральный угол, стягиваемый дугой, длина которой равна радиусу окружности; градусная мера такого угла содержит $y = \frac{180}{\pi} = 57,2957\ldots$ градусов. Для того чтобы от радианной меры угла x перейти к его градусной мере y, нужно величину x умножить на число $\frac{180}{\pi}$.

Радианная мера x некоторого угла равна также двойной площади A сектора, вырезаемого этим углом из круга единичного радиуса; в самом деле, эта площадь относится ко всей площади круга так, как длина дуги относится к длине всей окружности: $\frac{x}{2\pi} = \frac{A}{\pi}$; итак, $x = 2A$.

Будем впредь под углом x подразумевать угол, радианная мера которого есть x. Угол, градусное измерение которого равно x, будем в дальнейшем, чтобы устранить всякую неясность, обозначать через x°.

Позднее станет совершенно очевидным, насколько выгодно пользоваться радианным измерением при разного рода аналитических операциях. Однако следует признать, что для практического употребления оно скорее неудобно. В самом деле, так как π — иррациональное число, то, сколько раз мы ни откладывали бы по кругу единичный угол, т. е. угол с радианной мерой, равной 1, мы никогда не вернемся в начальную

точку. Обычное же измерение таково, что после откладывания 1 градуса 360 раз или 90 градусов 4 раза мы возвращаемся в исходную точку.

3. График функции. Обратные функции. Часто характер функции чрезвычайно ясно выражается с помощью простого графика. Если $(x,\, u)$ — координаты на плоскости относительно двух взаимно перпендикулярных осей, то линейные функции

$$u = ax + b$$

изображаются прямыми линиями; квадратические функции

$$u = ax^2 + bx + c$$

— параболами; функция

$$u = \frac{1}{x}$$

— гиперболой, и т. д. По определению, *график* некоторой функции $u = f(x)$ состоит из всех тех точек плоскости, координаты которых $(x,\, u)$ связаны уравнением $u = f(x)$. Функции $\sin x$, $\cos x$, $\operatorname{tg} x$ представлены графически на рис. 151 и 152. Эти графики наглядно показывают, как возрастают или убывают функции при изменении x.

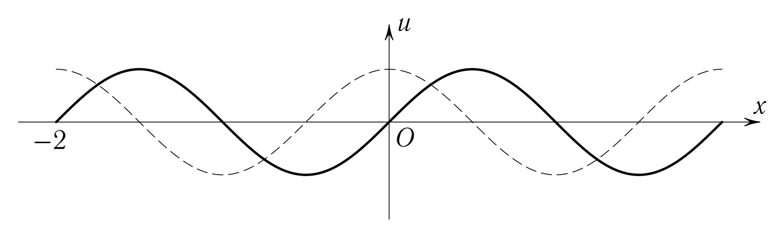


Рис. 151. Графики функций $\sin x$ и $\cos x$

Одним из важных методов, служащих для введения новых функций, является следующий. Исходя из некоторой известной функции $F(X)$, можно попытаться решить уравнение $U = F(X)$ относительно X — так, чтобы X было выражено как функция от U:

$$X = G(U).$$

Тогда функция $G(U)$ называется *обратной* относительно функции $F(X)$. Этот процесс приводит к результату однозначно только в том случае, если функция $U = F(X)$ определяет взаимно однозначное отображение области изменения X на область изменения U, т. е. если неравенство $X_1 \neq X_2$ всегда влечет за собой неравенство $F(X_1) \neq F(X_2)$. Только при этом условии каждому значению U будет соответствовать единственное значение X. Здесь будет кстати вспомнить приведенный выше пример, в котором роль независимого переменного X играл любой треугольник на плоскости, а в качестве функции $U =$

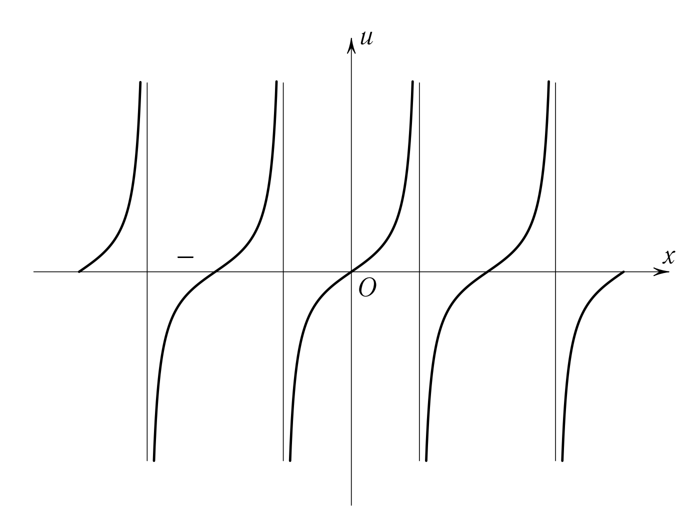


Рис. 152. $u = \operatorname{tg} x$

$F(X)$ рассматривался его периметр. Очевидно, что такое отображение множества S треугольников на множество T положительных чисел не является взаимно однозначным, так как имеется бесконечное количество различных треугольников с одним и тем же периметром. Итак, в этом случае соотношение $U = F(X)$ не может служить для однозначного определения обратной функции. С другой стороны, функция $m = 2n$, где n пробегает множество S всех целых чисел, а m — множество T четных чисел, напротив, дает взаимно однозначное соответствие между двумя множествами, и обратная функция $n = \frac{m}{2}$ будет определена. В качестве другого примера данного однозначного отображения приведем функцию

$$u = x^3.$$

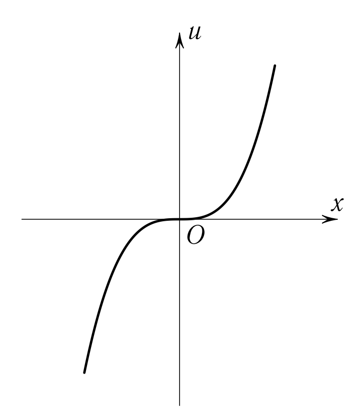


Рис. 153. $u = x^3$

Когда x пробегает множество всех действительных чисел, u тоже пробегает множество всех действительных чисел, принимая каждое значение один и только один раз. Однозначно определенная в этом примере обратная функция имеет вид

$$x = \sqrt[3]{u}.$$

В случае функции

$$u = x^2$$

обратная функция не определена однозначно. В самом деле, в силу того, что $u = x^2 = (-x)^2$, каждому положительному значению u соответствуют *два* разных значения («прообраза») x. Но если под символом $\sqrt{u}$ подразумевать (как это часто и делается) *положительное* число, квадрат которого есть x, то обратная функция

$$x = \sqrt{u}$$

существует, если только мы условимся, что будем рассматривать лишь положительные значения x и u.

Существование обратной функции может быть сразу установлено при взгляде на график данной функции. Обратная функция существует, определяясь однозначно, в том случае, если каждому значению u соответствует только одно значение x. Геометрически это означает, что нет такой прямой, параллельной оси x, которая пересекала бы график более чем в одной точке. Само собой разумеется, что так будет в том случае, если функция $u = f(x)$ монотонная, т. е. или все время возрастающая, или, наоборот, все время убывающая (при возрастании x). Например, если функция $u = f(x)$ всюду возрастающая, то при $x_1 < x_2$ мы всегда имеем $u_1 = f(x_1) < u_2 = f(x_2)$. Следовательно, для данного значения u существует не более одного значения x такого, что $u = f(x)$, и обратная функция будет определяться однозначно. График обратной функции $x = g(u)$ получается из данного графика просто симметрией относительно пунктирной прямой (рис. 154); при этом оси x и u меня-

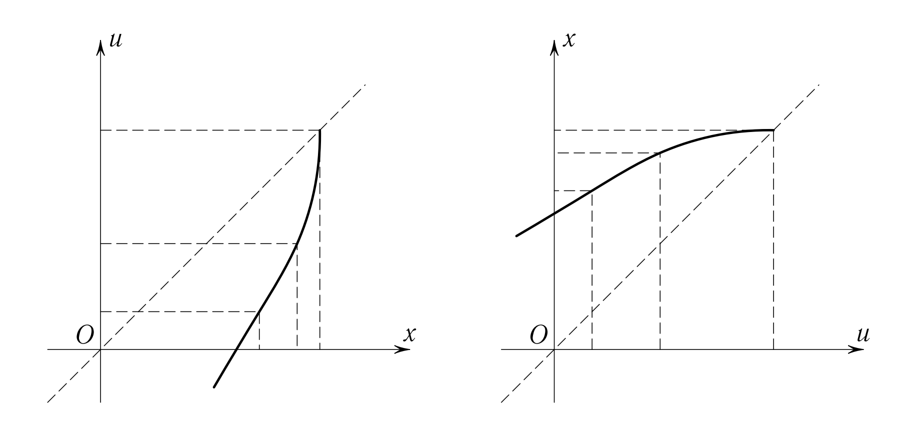


Рис. 154. Взаимно обратные функции

ются местами. Новое положение графика изображает x как функцию от u. В основном положении график указывает значение u как высоты над горизонтальной осью x, в то время как после симметрии вновь полученный график указывает значение x как высоты над горизонтальной

осью u.

Рассуждения этого параграфа можно иллюстрировать на примере функции

$$u = \operatorname{tg} x.$$

Эта функция *монотонна* в промежутке $-\frac{\pi}{2} < x < \frac{\pi}{2}$ (рис. 152): значения u, все время возрастающие вместе с x, изменяются от $-\infty$ до $+\infty$; отсюда ясно, что обратная функция

$$x = g(u)$$

определена для всех значений u. Эту функцию обозначают $\operatorname{arctg} u$. Таким образом, $\operatorname{arctg}(1) = \frac{\pi}{4}$, поскольку $\operatorname{tg}\frac{\pi}{4} = 1$. График $\operatorname{arctg} u$ изображен на рис. 155.

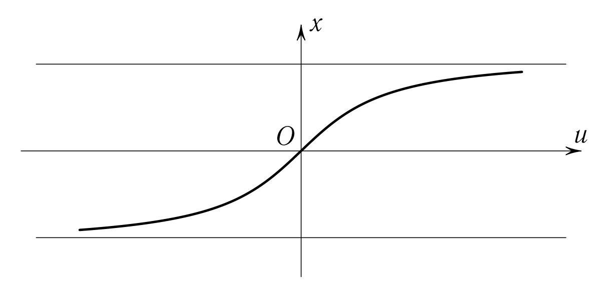


Рис. 155. $x = \operatorname{arctg} u$

4. Сложные функции. Вторым важным методом создания новых функций из двух или большего числа данных является *составление сложных функций* («*композиция*»). Так, например, функция

$$u = f(x) = \sqrt{1+x^2}$$

«составлена» из двух простых функций

$$z = g(x) = 1 + x^2, \quad u = h(z) = \sqrt{z}$$

и может быть записана так:

$$u = f(x) = h(g[x])$$

(читается «h от g от x»). Аналогично, функция

$$u = f(x) = \frac{1}{\sqrt{1-x^2}}$$

составлена из трех функций

$$z = g(x) = 1 - x^2, \quad w = h(z) = \sqrt{z}, \quad u = k(w) = \frac{1}{w},$$

так что можно написать

$$u = f(x) = k(h[g(x)]).$$

Функция

$$u = f(x) = \sin\frac{1}{x}$$

составлена из двух функций

$$z = g(x) = \frac{1}{x}, \qquad u = h(z) = \sin z.$$

Функция $f(x) = \sin\frac{1}{x}$ не определена при $x = 0$, так как при $x = 0$ выражение $\frac{1}{x}$ не имеет смысла. График этой замечательной функции находится в некоторой связи с графиком синуса. Мы знаем, что $\sin z = 0$ при $z = k\pi$, где k — произвольное положительное или отрицательное целое число. Кроме того,

$$\sin z = \begin{cases} 1 & \text{при } z = (4k+1)\frac{\pi}{2}, \\ -1 & \text{при } z = (4k-1)\frac{\pi}{2}, \end{cases}$$

где k — произвольное целое число. Отсюда имеем

$$\sin\frac{1}{x} = \begin{cases} 0 & \text{при } x = \frac{1}{k\pi}, \\ 1 & \text{при } x = \frac{2}{(4k+1)\pi}, \\ -1 & \text{при } x = \frac{2}{(4k-1)\pi}. \end{cases}$$

Если мы последовательно станем полагать $k = 1, 2, 3, 4, \ldots$, знаменатели этих дробей будут возрастать неограниченно и, следовательно, значения x, при которых функция $\sin\frac{1}{x}$ имеет значения $1, -1, 0$, будут сгущаться все больше и больше около точки $x = 0$. Между каждой такой точкой и началом будет всегда бесконечное количество колебаний. График этой функции показан на рис. 156.

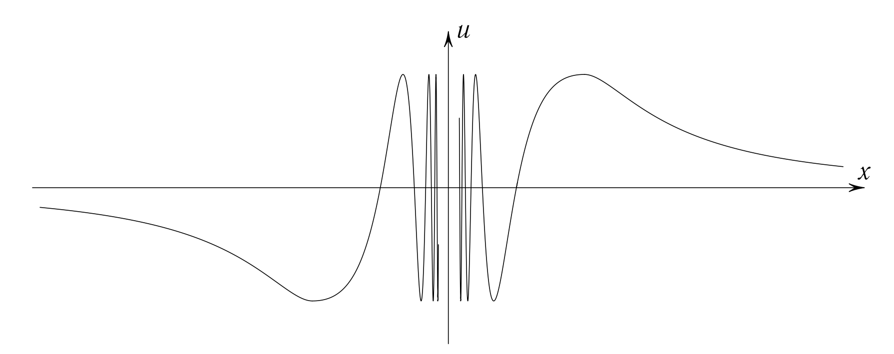


Рис. 156. $u = \sin\frac{1}{x}$

5. Непрерывность. Графики уже рассмотренных функций дают интуитивное представление о свойстве, называемом непрерывностью. Точное определение этого понятия мы дадим в § 4, после того как понятие предела будет поставлено на строго логический фундамент. Здесь

же, ограничиваясь описательной формулировкой, мы скажем, что функция непрерывна, если ее график есть плавная, нигде не «прерывающаяся» кривая. Чтобы уяснить себе, является ли функция $u = f(x)$ непрерывной *в точке* $x = x_1$, заставим независимую переменную x приближаться непрерывно справа и слева к значению x_1. При этом значения функции $u = f(x)$ меняются, если только эта функция не является постоянной в окрестности точки x_1. Если оказывается, что значение функции $f(x)$ неограниченно приближается к значению $f(x_1)$ этой функции в выбранной точке $x = x_1$ («стремится к пределу $f(x_1)$»), и притом *независимо от того*, приближается ли x к x_1 *с одной стороны или с другой*, то тогда говорят, что функция $f(x)$ *непрерывна в точке* x_1. Если это имеет место в каждой точке x_1 из некоторого интервала, то говорят, что функция *непрерывна в этом интервале*.

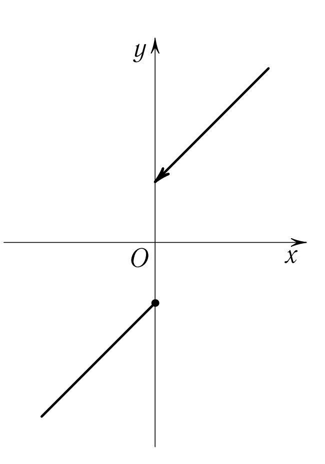


Рис. 157. Разрыв «скачком»

Хотя каждая функция, представляемая плавным графиком, непрерывна, очень легко определить и такие функции, которые не везде непрерывны. Например, функция на рис. 157, определенная для всех значений x с помощью формул

$$f(x) = 1 + x \qquad \text{при } x > 0,$$
$$f(x) = -1 + x \qquad \text{при } x \leqslant 0,$$

разрывна в точке $x_1 = 0$, в которой она имеет значение -1. Если мы станем чертить карандашом график этой функции, нам придется в этой точке оторвать карандаш от бумаги. Когда мы приближаемся к значению $x_1 = 0$ справа, то $f(x)$ стремится к $+1$. Но значение это отличается от значения функции в самой этой точке, именно -1.

Одно то обстоятельство, что функция $f(x)$ стремится к -1, когда x стремится к нулю *слева*, еще недостаточно для установления непрерывности.

Функция $f(x)$, определенная для всех значений x с помощью формул

$$f(x) = 0 \quad \text{при} \quad x \neq 0, \qquad f(0) = 1,$$

при $x_1 = 0$ имеет разрыв другого вида. Здесь существуют пределы и справа и слева, и они равны между собой, но это общее предельное значение отлично от $f(0)$. Еще иного типа разрыв дается функцией, график которой изображен на рис. 158,

$$u = f(x) = \frac{1}{x^2}$$

в точке $x = 0$. Если мы заставим x стремиться к 0 с любой стороны, то u неизменно будет стремиться к бесконечности, но график функции «прерывается» в этой точке, причем малым изменениям независимого переменного x в окрестности точки $x = 0$ могут соответствовать очень большие изменения зависимого переменного u. Строго говоря, значение функции не определено при $x = 0$, поскольку мы не считаем бесконечность числом, и поэтому нельзя говорить, что функция $f(x)$ *равна* бесконечности при $x = 0$. Итак, мы говорим только, что функция $f(x)$ «стремится к бесконечности», когда x приближается к нулю.

Совсем иной характер разрыва, наконец, у функции $u = \sin\frac{1}{x}$ в точке $x = 0$ (рис. 156).

Приведенные примеры показывают несколько различных типических случаев, когда функция перестает быть непрерывной в некоторой точке $x = x_1$.

1) Может случиться, что функция станет непрерывной в точке $x = x_1$ после того, как надлежащим образом будет определено или будет изменено уже определенное значение ее при $x = x_1$. Например, функция $u = \frac{x}{x}$ постоянно равна 1 при $x \neq 0$; она не определена при $x = 0$, поскольку $\frac{0}{0}$ — лишенный смысла символ. Но если в этом примере мы условимся считать, что значение $u = 1$ соответствует также и значению $x = 0$, то функция, «расширенная» таким образом, становится непрерывной во всех точках без исключения. Тот же результат будет достигнут, если мы изменим значение функции при $x = 0$ во втором из приведенных выше примеров, и вместо $f(0) = 1$ положим $f(0) = 0$. Разрывы этого рода называются *устранимыми*.

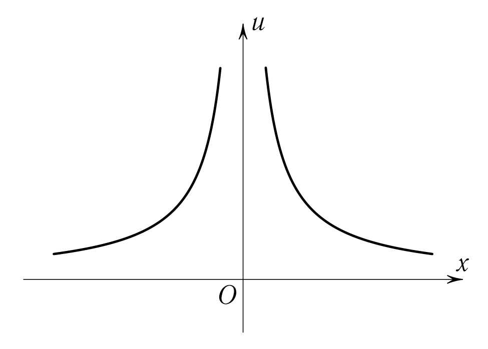


Рис. 158. Разрыв с уходом в бесконечность

2) Функция стремится к различным пределам в зависимости от того, справа или слева x приближается к x_1, как на рис. 157.

3) Не существует предела ни с одной, ни с другой стороны, как на рис. 156.

4) Функция стремится к бесконечности, когда x приближается к x_1 (рис. 158).

Разрывы трех последних типов называются *существенными* или *неустранимыми*, они не могут быть устранены с помощью надлежащего определения значения функции в одной лишь точке $x = x_1$.

Упражнения. 1) Наметьте графики функций $\frac{x-1}{x^3}$, $\frac{x^2-1}{x^2+1}$, $\frac{x}{(x^2-1)(x^2+1)}$ и найдите точки разрыва.

2) Наметьте графики функций $x\sin\frac{1}{x}$ и $x^2\sin\frac{1}{x}$; проверьте, что непрерывность не нарушена в точке $x=0$, если принять, что $u=0$ при $x=0$ в обоих случаях.

*3) Покажите, что функция $\operatorname{arctg}\frac{1}{x}$ имеет разрыв второго типа (скачок) при $x=0$.

***6. Функции нескольких переменных.** Вернемся к систематическому рассмотрению понятия функции. Если независимым переменным P является точка плоскости с координатами x, y и если каждой такой точке P соответствует единственное число u (например, u может быть расстоянием точки P от начала), тогда принято писать

$$u = f(x,\, y).$$

Это обозначение употребляется также и в том случае, если, как это часто бывает, две величины x и y явно указываются самими условиями задачи как независимые переменные. Например, давление u газа есть функция объема x и температуры y; площадь u треугольника есть функция $u = f(x,\, y,\, z)$ длин трех его сторон x, y, z.

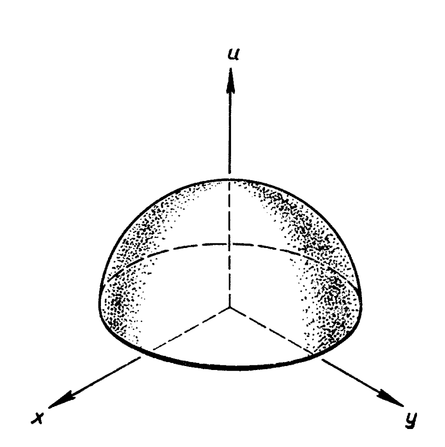


Рис. 159. Полусфера

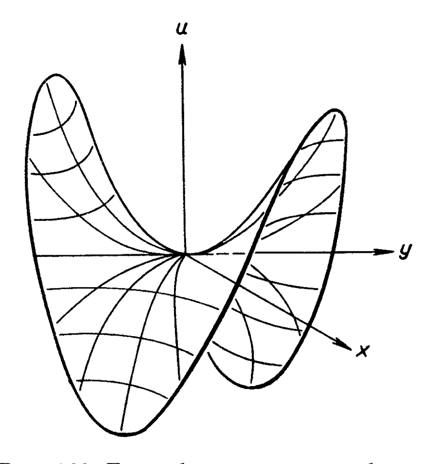


Рис. 160. Гиперболический параболоид

Так же как график дает геометрическое представление функции одного переменного, можно получить и геометрическое представление функции $u = f(x,\, y)$ двух переменных в виде поверхности в трехмерном пространстве с переменными x, y, u в качестве координат. Каждой точке x, y в плоскости x, y мы сопоставляем точку пространства с координатами x, y и $u = f(x,\, y)$. Таким образом, $u = \sqrt{1-x^2-y^2}$ представляется поверхностью сферы с уравнением $u^2+x^2+y^2=1$,

линейная функция $u = ax + by + c$ — плоскостью, функция $u = xy$ — гиперболическим параболоидом, и т. д.

Можно дать и другое представление функции $u = f(x, y)$, притом не выходя за пределы плоскости x, y, именно с помощью *линий уровня* (*горизонталей*). Вместо того чтобы рассматривать трехмерный «ландшафт» поверхности $u = f(x, y)$ в трехмерном пространстве, мы вычерчиваем, как это иногда делают на географических картах, «линии уровня» функции, являющиеся проекциями на плоскость x, y всех точек поверхности, находящихся на одном и том же расстоянии u по вертикали от плоскости x, y. Эти линии уровня имеют уравнения вида $f(x, y) = c$, где c постоянно для каждой кривой. Так, например, функция $u = x + y$ характеризуется рис. 163. Линии уровня поверхности сферы представляют собой семейство концентрических окружностей. Функция $u = x^2 + y^2$, которой соответствует параболоид вращения, характеризуется также окружностями (рис. 165). Числами, отнесенными к каждой кривой, можно указывать высоту $u = c$.

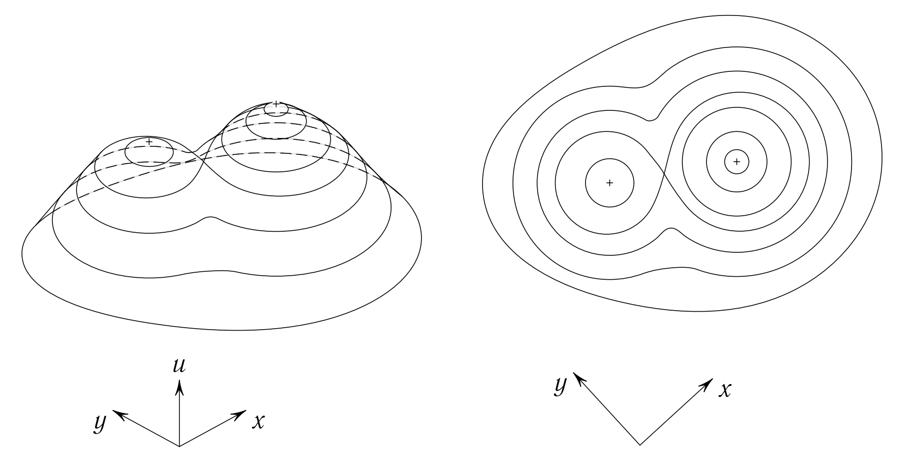


Рис. 161. Поверхность вида $u = f(x, y)$

Рис. 162. Линии уровня поверхности, изображенной на рис. 161

Функции нескольких переменных встречаются в физике при описании движения непрерывной среды или каких угодно протяженных объектов. Рассмотрим хотя бы струну, натянутую между двумя точками на оси x и затем деформированную таким образом, что частица с координатой x отодвинута на некоторое определенное расстояние перпендикулярно к оси. Если струна будет отпущена, то она придет в движение, т. е. начнет колебаться; тогда точка (частица) струны с начальной координатой x в момент времени t будет находиться на расстоянии $u = f(x, t)$ от

оси x. Движение струны будет полностью определено, если только будет известна функция $u = f(x, t)$.

Определение непрерывности, данное для функций одного переменного, распространяется непосредственно и на функции нескольких переменных. Говорят, что функция $u = f(x, y)$ непрерывна в точке $x = x_1$, $y = y_1$, если значение $f(x, y)$ всегда стремится к значению $f(x_1, y_1)$, когда точка x, y приближается к точке x_1, y_1 по любому направлению или любым способом.

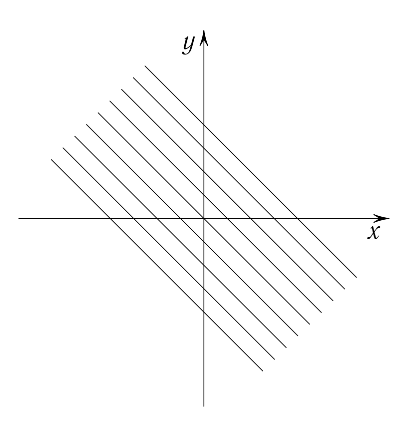


Рис. 163. Линии уровня поверхности $u = x + y$

Впрочем, имеется одно существенное различие между функциями одного и нескольких переменных. В последнем случае понятие обратной функции теряет смысл, так как мы не можем решить уравнение $u = f(x, y)$, например $u = x + y$, так, чтобы *каждое* из независимых переменных x и y было бы выражено с помощью только одного переменного u. Но это различие между функциями одного и нескольких переменных исчезает, если мы перейдем, далее, к рассмотрению преобразований или отображений.

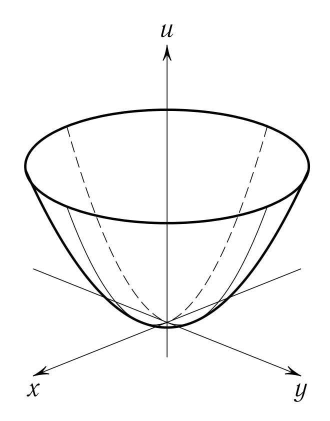


Рис. 164. Параболоид вращения

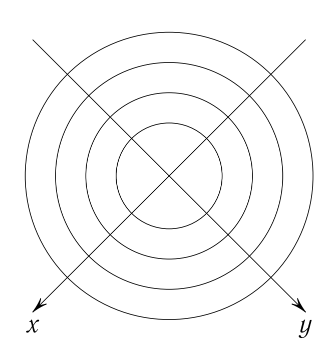


Рис. 165. Соответствующие линии уровня

***7. Функции и преобразования.** Соответствие между точками некоторой прямой l, характеризуемыми координатой x на этой прямой, и точками некоторой другой прямой l', характеризуемыми координатой x', есть не что иное, как некоторая функция $x' = f(x)$. В случае взаимно однозначного соответствия имеем также и обратную функцию $x = g(x')$. Простейшим примером является проективное преобразование, которое задается в самом общем случае дробно-линейной функцией вида

$$x' = f(x) = \frac{ax+b}{cx+d},$$

где a, b, c, d — постоянные (мы это утверждаем здесь без доказательства).

В этом примере обратная функция имеет вид $x = \dfrac{-dx'+b}{cx'-a}$.

В случае, если устанавливается отображение плоскости π с координатной системой x, y на другую плоскость π' с координатной системой x', y', соотношение между точками не может быть задано одной функцией $x' = f(x)$; здесь приходится иметь дело с двумя функциями двух переменных

$$x' = f(x,\, y),$$
$$y' = g(x,\, y).$$

Например, проективное преобразование задается системой функций

$$x' = \frac{ax+by+c}{gx+hy+k},$$
$$y' = \frac{dx+ey+f}{gx+hy+k},$$

где a, b, $\ldots$, k — постоянные, а x, y и x', y', как сказано, — соответственные координаты в двух плоскостях. Теперь идея обратного отображения снова приобретает смысл. Мы просто должны *решить данную систему* уравнений относительно x и y, выразив их через x' и y'. Геометрически это сводится к осуществлению обратного отображения плоскости π' на плоскость π. Это отображение будет однозначно определено, если соответствие между точками обеих плоскостей взаимно однозначное.

Преобразования плоскости, изучаемые в топологии, задаются не простыми алгебраическими уравнениями, а произвольной системой двух функций

$$x' = f(x,\, y),$$
$$y' = g(x,\, y),$$

при условии, чтобы ими определялось взаимно однозначное и взаимно непрерывное преобразование.

Упражнения. *1) Покажите, что преобразование инверсии (стр. 162–165) в единичном круге аналитически задается уравнениями $x' = \dfrac{x}{x^2+y^2}$, $y' = \dfrac{y}{x^2+y^2}$. Найдите обратное преобразование. Докажите аналитически,

что путем инверсии совокупность прямых и окружностей преобразуется в совокупность опять-таки прямых и окружностей.

2) Докажите, что преобразованием $x' = \dfrac{ax+b}{cx+d}$ четверка точек на оси x переводится в четверку точек на оси x с тем же двойным отношением (см. стр. 216).

§ 2. Пределы

1. Предел последовательности a_n. Определение непрерывности функции, как мы это уже видели в § 1, основывается на понятии предела. До сих пор мы пользовались этим понятием в более или менее интуитивной форме. В настоящем и последующих разделах мы введем его более систематическим путем. Поскольку последовательности несколько проще, чем функции непрерывного переменного, мы начнем с изучения последовательностей.

В главе II мы имели дело с числовыми последовательностями a_n и изучали их пределы при условии, что n неограниченно увеличивается или «стремится» к бесконечности. Например, последовательность с общим членом $a_n = \dfrac{1}{n}$:

$$1,\ \frac{1}{2},\ \frac{1}{3},\ \ldots,\ \frac{1}{n},\ \ldots \tag{1}$$

при неограниченном возрастании n имеет предел 0:

$$\frac{1}{n} \to 0 \quad \text{при} \quad n \to \infty. \tag{2}$$

Постараемся выразить точно, что под этим подразумевается. При продвижении по последовательности все дальше и дальше мы видим, что члены становятся все меньше и меньше. После 100-го члена члены уже меньше $\dfrac{1}{100}$, после 1000-го — меньше $\dfrac{1}{1000}$, и т. д. Ни один из членов не равен в действительности 0. Но если мы продвинемся *достаточно далеко* по последовательности (1), то мы можем быть уверены, что каждый из ее членов будет отличаться от 0 *сколь угодно мало.*

В этом объяснении может смущать единственно то, что смысл подчеркнутых слов не вполне ясен. Что значит «достаточно далеко» и что значит «сколь угодно мало»? Если мы сумеем придать точный смысл этим выражениям, то этим самым будет установлен точный смысл понятия предела последовательности.

Геометрическая интерпретация поможет нам разобраться в интересующем нас вопросе. Если мы изобразим члены последовательности (1) в виде соответствующих им точек на числовой оси, то заметим, что члены последовательности в нашем примере «накопляются» или «сгущаются» около точки 0. Выберем на числовой оси некоторый интервал I с центром в точке 0 и общей длиной 2ε, так чтобы интервал простирался на

расстояние ε с каждой стороны от точки 0. Если мы возьмем $\varepsilon = 10$, то, конечно, *все* члены $a_n = \frac{1}{n}$ нашей последовательности будут лежать внутри интервала I. Если же мы возьмем $\varepsilon = \frac{1}{10}$, то несколько первых членов окажутся лежащими вне интервала I; однако все члены, начиная с a_{11}, а именно

$$\frac{1}{11},\ \frac{1}{12},\ \frac{1}{13},\ \frac{1}{14},\ \ldots,$$

будут лежать внутри I. Даже при $\varepsilon = \frac{1}{1000}$ лишь первая тысяча членов последовательности не попадет внутрь интервала I, тогда как бесконечное множество членов, начиная с a_{1001},

$$a_{1001},\ a_{1002},\ a_{1003},\ \ldots$$

окажется внутри него. Ясно, что это рассуждение справедливо для любого положительного числа ε: если положительное ε выбрано, то независимо от того, как оно мало, мы всегда можем подобрать настолько большое целое число N, что

$$\frac{1}{N} < \varepsilon.$$

Отсюда следует, что все члены a_n последовательности, для которых $n \geqslant N$, будут лежать внутри интервала I, и только конечное число членов a_1, a_2, …, a_{N-1} может лежать вне его. Основные моменты здесь таковы: *во-первых*, длина интервала I определяется произвольно посредством выбора ε. *Затем* может быть подобрано подходящее целое число N. Этот процесс первоначального выбора числа ε и последующего подбора целого числа N может быть осуществлен при любом положительном ε независимо от его малости; тем самым устанавливается точный смысл утверждения, что «все члены последовательности (1) отличаются от 0 сколь угодно мало, если только мы достаточно далеко продвинемся по последовательности».

Подведем итоги: пусть ε — какое-нибудь положительное число. Тогда мы можем подобрать такое целое положительное число N, что все члены a_n последовательности (1), для которых $n \geqslant N$, будут лежать внутри интервала I длины 2ε с центром в точке ε. Таков смысл предельного соотношения (2).

Опираясь на этот пример, мы можем теперь дать точное определение следующего общего утверждения: «Последовательность действительных чисел a_1, a_2, a_3, … имеет предел a». Число a мы заключаем внутрь некоторого интервала I числовой оси: если этот интервал мал, то некоторые числа a_n могут лежать вне этого интервала, но как только n становится достаточно большим, скажем, бо́льшим или равным некоторому числу N, то все числа a_n, для которых $n \geqslant N$, должны лежать внутри

интервала I. Конечно, может случиться, что придется брать очень большое целое число N, если интервал I выбран очень малым; однако, как бы мал ни был этот интервал I, такое целое число N должно существовать, раз предполагается, что последовательность имеет предел.

Тот факт, что последовательность a_n имеет предел a, выражается символически следующей записью:

$$\lim a_n = a \quad \text{при } n \to \infty, \qquad \lim_{n\to\infty} a_n = a,$$

или просто

$$a_n \to a \quad \text{при } n \to \infty$$

(«предел a_n равен a», или «a_n стремится к пределу a»). Если последовательность имеет предел в указанном смысле, она называется *сходящейся.* Определение сходимости последовательности a_n можно сформулировать более сжато, а именно следующим образом:

Последовательность a_1, a_2, a_3, ... *имеет предел* a *при неограниченном возрастании* n, *если каждому сколь угодно малому положительному числу* ε *можно поставить в соответствие такое целое положительное число* N (*зависящее от* ε), *что неравенство*

$$|a - a_n| < \varepsilon \tag{3}$$

выполняется для всех значений

$$n \geqslant N.$$

Такова общая, «абстрактная» формулировка понятия предела последовательности. Немудрено, если тот, кто встречается с ней впервые, не может сразу схватить и исчерпать ее содержание. К несчастью, авторы некоторых руководств, стоящие на позиции, граничащей со снобизмом, преподносят читателю это определение без тщательной подготовки, как будто снизойти до разъяснений ниже достоинства математика.

Определение предполагает своего рода дискуссию между двумя лицами A и B. A выдвигает требование: заданная величина a должна быть приближена числами a_n так, чтобы ошибка не превышала границы $\varepsilon = \varepsilon_1$; B отвечает на это требование указанием, что существует такое целое $N = N_1$, что все члены a_n, следующие за a_{N_1}, удовлетворяют этому условию. Тогда A становится более требовательным и предлагает новую, меньшую границу $\varepsilon = \varepsilon_2$; B снова встречает это требование подбором некоторого, может быть значительно большего, целого числа $N = N_2$, обладающего аналогичным свойством, и т. д. *Если B может удовлетворить A независимо от того, какую малую границу назначает A, то мы имеем дело с положением, которое кратко выражается соотношением: $a_n \to a$.*

Имеется определенная психологическая трудность в том, чтобы составить правильное представление о понятии предела. Наша интуиция

предполагает «динамическую» идею предела как результат процесса «движения»: мы продвигаемся по ряду целых чисел $1, 2, 3, \ldots, n, \ldots$ и при этом наблюдаем за поведением последовательности a_n. Мы ждем, что числа a_n должны все меньше и меньше отличаться от числа a. Но эта «естественная» точка зрения не поддается ясной математической формулировке. Чтобы прийти к точному определению, надо обратить ход рассуждения: вместо того чтобы прежде всего обращаться к независимому переменному n, а затем уже к переменному a_n, мы должны основывать наше определение на том, что следует делать, если мы по существу хотим проконтролировать утверждение $a_n \to a$. При такой постановке вопроса мы прежде всего должны выбрать произвольно малый интервал около a и затем решить: можем ли мы добиться, чтобы a_n с помощью выбора достаточно большого n в него попало. Затем, вводя символы ε и N для обозначения «произвольно малого интервала» и «достаточно большого n», мы приходим к точному определению предела.

Обращаясь теперь к другому примеру, рассмотрим последовательность

$$\frac{1}{2},\ \frac{2}{3},\ \frac{3}{4},\ \frac{4}{5},\ \ldots,\ \frac{n}{n+1},\ \ldots,$$

где $a_n = \dfrac{n}{n+1}$. Я утверждаю, что $\lim a_n = 1$. Если вы выберете интервал с центром в точке 1 и возьмете $\varepsilon = \dfrac{1}{10}$, то я смогу удовлетворить вашему требованию (3), выбрав $N = 10$; в самом деле,

$$0 < 1 - \frac{n}{n+1} = \frac{n+1-n}{n+1} = \frac{1}{n+1} < \frac{1}{10}$$

при $n \geqslant 10$. Если вы усилите ваше требование, выбирая $\varepsilon = \dfrac{1}{1000}$, то я снова могу ему удовлетворить, выбирая $N = 1000$; и так далее — для любого положительного числа ε, которое вы пожелаете выбрать, независимо от его малости: действительно, мне только нужно будет выбрать любое целое N, большее чем $\dfrac{1}{\varepsilon}$. Этот процесс, заключающийся, во-первых, в выборе произвольно малого интервала длины 2ε вокруг числа a и, во-вторых, в доказательстве того, что все члены последовательности a_n находятся на расстоянии, меньшем чем ε от a, раз только мы продвинемся достаточно далеко по последовательности, и есть не что иное, как подробное описание того факта, что $a_n \to a$. Если члены последовательности $a_1, a_2, a_3, \ldots$ представлены в виде бесконечных десятичных дробей, то утверждение

$$\lim a_n = a$$

обозначает попросту то, что для любого целого положительного m первые m цифр числа a_n совпадают с первыми m цифрами бесконечного десятичного разложения числа a, раз только n выбрано достаточно боль-

шим, скажем, бóльшим некоторого значения N (зависящего от m).[1] Это просто соответствует выбору ε в форме 10^{-m}.

Существует другое, и очень выразительное, определение понятия предела. Если $\lim a_n = a$ и если мы заключим число a в интервал I, то, независимо от малости интервала I все числа a_n при n, бóльшем некоторого целого числа N, будут лежать в интервале I, так что не больше чем конечное число $N-1$ членов из числа следующих:

$$a_1,\ a_2,\ a_3,\ \ldots,\ a_{N-1},$$

могут лежать вне интервала I. Если интервал I очень мал, то число N может быть очень большим, скажем, равным 100 или даже 1000 биллионам, и все же лишь конечное число членов последовательности будет лежать вне интервала I, в то время как бесконечное множество оставшихся членов попадет в интервал I. Можно условиться говорить о членах некоторой бесконечной последовательности, что «почти все» они обладают некоторым свойством, если лишь конечное число их (неважно, как оно будет велико) не обладает этим свойством. Так, например, «почти все» целые положительные числа больше 1 000 000 000 000. Используя эту терминологию, мы видим, что утверждение

$$\lim a_n = a$$

эквивалентно следующему утверждению: *если I есть любой интервал с центром в точке a, то почти все числа a_n лежат в этом интервале.*

Не мешает мимоходом отметить, что нет необходимости предполагать, что все члены a_n последовательности непременно имеют различные значения. В частности, совершенно допустимо, чтобы несколько из них, или бесконечное число, или даже, наконец, все числа a_n были равны предельному значению a. Например, вполне законно рассматривать последовательность $a_1 = 0$, $a_2 = 0$, $\ldots$, $a_n = 0$, $\ldots$, и ее предел есть, конечно, 0.

Как уже указано, последовательность a_n с пределом a называется *сходящейся*. Последовательность, не имеющая предела, называется *расходящейся*.

Упражнения. Докажите: 1) Последовательность с общим членом $a_n = \dfrac{n}{n^2+1}$ имеет предел, равный 0. $\left(\text{\textit{Указание}: } a_n = \dfrac{1}{n+\frac{1}{n}} \text{ меньше } \dfrac{1}{n} \text{ и больше } 0.\right)$

2) Последовательность $a_n = \dfrac{n^2+1}{n^3+1}$ имеет предел 0. $\left(\text{\textit{Указание}: } a_n = \dfrac{1+\frac{1}{n^2}}{n+\frac{1}{n^2}} \text{ лежит между 0 и } \dfrac{2}{n}.\right)$

[1] Это утверждение не вполне точно. — *Прим. ред. наст. изд.*

3) Последовательность 1, 2, 3, 4, ..., а также «колеблющиеся» последовательности

$$1,\ 2,\ 1,\ 2,\ 1,\ 2,\ \ldots,$$

$$-1,\ 1,\ -1,\ 1,\ -1,\ 1,\ \ldots \quad (\text{т. е. } a_n = (-1)^n),$$

и

$$1,\ \frac{1}{2},\ 1,\ \frac{1}{3},\ 1,\ \frac{1}{4},\ 1,\ \frac{1}{5},\ \ldots$$

не имеют пределов.

Если в последовательности a_n члены возрастают таким образом, что рано или поздно становятся больше любого наперед назначенного числа K, то принято говорить, что a_n *стремится к бесконечности*, и тогда пишут: $\lim a_n = \infty$ или $a_n \to \infty$. Например, $n^2 \to \infty$ и $2^n \to \infty$. Эта терминология удобна, хотя и не вполне последовательна, так как символ ∞ не принято рассматривать как число. *Последовательность, стремящаяся к бесконечности, считается расходящейся.*

Упражнение. Докажите, что последовательность $a_n = \dfrac{n^2+1}{n}$ стремится к
бесконечности; аналогично в случае

$$a_n = \frac{n^2+1}{n+1}, \qquad a_n = \frac{n^3-1}{n+1}, \qquad a_n = \frac{n^n}{n^2+1}.$$

Начинающие часто впадают в ошибку, думая, что переход к пределу при $n \to \infty$ может быть выполнен совершенно просто путем подстановки $n = \infty$ в выражение общего члена a_n. Например, $\dfrac{1}{n} \to 0$ потому, что $\dfrac{1}{\infty} = 0$. Но символ ∞ не является числом, и его употребление в выражении $\dfrac{1}{\infty}$ незаконно. Попытка представить себе предел последовательности как «последний» член последовательности a_n при $n = \infty$ не попадает в цель и затемняет правильное понимание существа дела.

2. Монотонные последовательности. В общем определении сходимости и предела на стр. 313 не содержится требований, так или иначе стесняющих характер приближения сходящейся последовательности a_1, a_2, a_3, ... к своему пределу a. Простейший тип приближения осуществляется так называемыми *монотонными* последовательностями, примером которых может служить следующая:

$$\frac{1}{2},\ \frac{2}{3},\ \frac{3}{4},\ \ldots,\ \frac{n}{n+1},\ \ldots$$

Каждый член этой последовательности больше предыдущего. Действительно,

$$a_{n+1} = \frac{n+1}{n+2} = 1 - \frac{1}{n+2} > 1 - \frac{1}{n+1} = \frac{n}{n+1} = a_n.$$

Последовательность такого рода, в которой $a_{n+1} > a_n$, называется *монотонно возрастающей*. Аналогично, последовательность, для которой

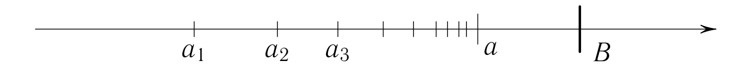


Рис. 166. Монотонная ограниченная поверхность

$a_n > a_{n+1}$, например такая, как $1, \frac{1}{2}, \frac{1}{3}, \ldots$, называется *монотонно убывающей.* Последовательности этих двух типов могут приближаться к своему пределу лишь с одной стороны: «слева» или «справа». В противоположность этому существуют последовательности колеблющиеся, например, $-1, \frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, \ldots$ Эта последовательность приближается к своему пределу 0 «с обеих сторон» (см. рис. 11, стр. 89).

Монотонные последовательности обладают особенно простыми свойствами. Такого рода последовательность может не иметь предела и возрастать неограниченно, подобно последовательности

$$1,\ 2,\ 3,\ 4,\ \ldots,$$

где $a_n = n$, или последовательности

$$2,\ 3,\ 5,\ 7,\ 11,\ 13,\ \ldots,$$

для которой a_n есть n-е простое число p_n. В этом случае последовательность стремится к бесконечности. Но если члены монотонно возрастающей последовательности остаются ограниченными, т. е. если каждый член меньше некоторой верхней границы B, заранее известной, то интуитивно ясно, что последовательность должна стремиться к некоторому определенному пределу a, не превышающему числа B. Мы сформулируем это положение как *принцип монотонных последовательностей*: *монотонно возрастающая последовательность, ограниченная сверху, сходится к некоторому пределу.* (Аналогичное утверждение имеет место относительно монотонно убывающей последовательности, ограниченной снизу.) Замечательно то, что значение предела a не должно быть известным заранее; теорема утверждает, что при выполнении указанных условий предел *существует.* Конечно, эта теорема справедлива лишь при условии, что предварительно введены иррациональные числа, и в противном случае не всегда бы оправдывалась: в самом деле, в главе II мы видели, что каждое иррациональное число (например, $\sqrt{2}$) является пределом монотонно возрастающей и ограниченной последовательности рациональных десятичных дробей, возникающих при обрывании некоторой бесконечной десятичной дроби на n-й цифре.

* Хотя принцип монотонных последовательностей интуитивно вполне очевиден и выглядит как абсолютная истина, однако не мешает, и даже весьма полезно, привести его вполне строгое доказательство в современном стиле. Чтобы это сделать, надо показать, что этот принцип есть логический вывод из определения действительного числа и определения предела.

Предположим, что числа a_1, a_2, a_3, ... образуют монотонно возрастающую, но ограниченную последовательность. Мы можем представить члены этой последовательности как бесконечные десятичные дроби

$$a_1 = A_1{,}p_1p_2p_3\ldots$$
$$a_2 = A_2{,}q_1q_2q_3\ldots$$
$$a_3 = A_3{,}r_1r_2r_3\ldots$$
$$\ldots\ldots\ldots\ldots\ldots$$

где A_i — целые числа, а p_i, q_i, r_i и т. д. — цифры от 0 до 9. Пробежим теперь вниз по столбцу целых чисел A_1, A_2, A_3, ... Так как последовательность a_1, a_2, a_3, ... *ограниченна*, то эти целые числа не могут возрастать бесконечно, а поскольку она *монотонно возрастает*, то целые числа последовательности A_1, A_2, A_3, ... *после достижения некоторого максимального значения должны стать постоянными*. Обозначим это максимальное значение символом A и предположим, что оно достигнуто в N_0-й строке. Станем теперь пробегать второй столбец p_1, q_1, r_1, ..., сосредоточивая свое внимание на членах N_0-й и последующих строк. Если x_1 есть наибольшая из цифр, появившаяся в этом столбце после N_0-й строки, то эта цифра будет появляться *всегда* после своего первого появления, которое, предположим, произошло в N_1-й строке, где $N_1 \geqslant N_0$. (Если бы цифра в этом столбце уменьшилась когда-либо впоследствии, то последовательность a_0, a_1, a_2, ... не была бы монотонно возрастающей.) Затем мы рассмотрим цифры p_2, q_2, r_2, ... третьего столбца. Рассуждение, подобное предыдущему, показывает, что начиная с некоторого целого числа $N_2 \geqslant N_1$ цифры третьего столбца постоянно равны некоторому числу x_2. Если мы повторим этот процесс для 4-го, 5-го, ... столбцов, то получим цифры x_3, x_4, x_5, ... и соответствующие целые числа N_3, N_4, N_5, ... Легко убедиться, что число

$$a = A_1{,}x_1x_2x_3x_4\ldots$$

есть предел последовательности a_1, a_2, a_3, ... В самом деле, пусть выбрано $\varepsilon \geqslant \geqslant 10^{-m}$; тогда для всех $n \geqslant N_m$ целая часть и первые m цифр после запятой в числах a_n и a будут совпадать между собой, так что разность $|a_n - a|$ не может превышать 10^{-m}. Так как это можно сделать для любого ε, как бы мало оно ни было, с помощью выбора достаточно большого m, то теорема доказана.

Эту теорему можно также доказать, основываясь на любом из данных в главе II определений действительного числа, например, взяв определение с помощью вложенных интервалов или дедекиндовых сечений. Такие доказательства можно найти в любом курсе анализа.

Принцип монотонных последовательностей мог бы быть применен в главе II при определении суммы и произведения двух положительных бесконечных десятичных дробей:

$$a = A{,}a_1a_2a_3\ldots$$
$$b = B{,}b_1b_2b_3\ldots$$

Два таких выражения не могут быть ни сложены, ни перемножены обычным путем, начиная с правого конца, поскольку нет никакого правого конца. (В качестве примера читатель может попытаться сложить следующие две бесконечные десятичные дроби: 0,333333... и 0,989898....) Но если символ x_n

обозначает конечную десятичную дробь, полученную в результате сложения конечных десятичных дробей, возникающих при «обрывании» на n-й цифре десятичных разложений a и b, то последовательность $x_1, x_2, x_3, \ldots$ будет монотонно возрастающей и ограниченной (границей может служить, например, число $A + B + 2$). Отсюда следует, что последовательность $x_1, x_2, x_3, \ldots$ имеет предел, и мы можем принять следующее определение:

$$a + b = \lim x_n.$$

Посредством подобного же процесса можно определить и произведение ab. Определения эти можно распространить и на все случаи, когда a и b — какие угодно положительные или отрицательные числа, применяя обычные правила арифметики.

Упражнение. Докажите, что суммой двух вышеуказанных бесконечных десятичных дробей является действительное число

$$1{,}323232\ldots = \frac{131}{99}.$$

Важность понятия предела в математике заключается в том, что *многие числа могут быть определены лишь как пределы* (часто как пределы монотонно возрастающих последовательностей). Вот почему поле рациональных чисел, в котором такие пределы могут не существовать, слишком узко для надобностей математики.

3. Число Эйлера e. Число e заняло видное место в математике рядом с архимедовым числом π сразу после опубликования Эйлером в 1748 г. сочинения «Introductio in Analysin Infinitorum». Это число является прекрасной иллюстрацией того, как принцип монотонных последовательностей может служить для определения нового действительного числа. Пользуясь обычной сокращенной записью для произведения n первых целых чисел

$$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n,$$

рассмотрим последовательность $a_1, a_2, a_3, \ldots$, где

$$a_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots + \frac{1}{n!}. \tag{4}$$

Члены последовательности a_n монотонно возрастают, поскольку a_{n+1} получается из a_n посредством прибавления положительного слагаемого $\frac{1}{(n+1)!}$. Кроме того, значения a_n ограничены сверху:

$$a_n < B = 3. \tag{5}$$

В самом деле, мы имеем

$$\frac{1}{s!} = \frac{1}{2} \cdot \frac{1}{3} \ldots \frac{1}{s} < \frac{1}{2} \cdot \frac{1}{2} \ldots \frac{1}{2} = \frac{1}{2^{s-1}};$$

отсюда вытекает, что

$$a_n < 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots + \frac{1}{2^{n-1}} = \\ = 1 + \frac{1 - \left(\frac{1}{2}\right)^n}{1 - \frac{1}{2}} = 1 + 2\left(1 - \left(\frac{1}{2}\right)^n\right) < 3,$$

причем мы использовали формулу стр. 85 для суммы n первых членов геометрической прогрессии. Но в таком случае в силу принципа монотонных последовательностей a_n должно стремиться к некоторому пределу при стремлении n к бесконечности; *этот предел обозначается буквой* e. Чтобы выразить тот факт, что $e = \lim a_n$, мы можем записать e в виде «бесконечного ряда»

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{n!} + \ldots \qquad (6)$$

Это «тождество» с рядом точек на конце есть просто другой способ для выражения двух следующих утверждений:

$$a_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{n!},$$
$$a_n \to e \quad \text{при} \quad n \to \infty.$$

Ряд (6) позволяет вычислить e с любой степенью точности. Например, сумма (с девятью цифрами) членов ряда (6) до $\frac{1}{12!}$ включительно равна числу

$$\sum = 2{,}71828182\ldots$$

(Проверьте!) «Ошибка», т. е. разность между этим приближенным и истинным значением e, может быть легко оценена. Для разности $e - \sum$ мы имеем выражение

$$\frac{1}{13!} + \frac{1}{14!} + \ldots < \frac{1}{13!}\left(1 + \frac{1}{13} + \frac{1}{13^2} + \frac{1}{13^3} + \ldots\right) = \frac{1}{13!}\,\frac{1}{1 - \frac{1}{13}} = \frac{1}{12} \cdot \frac{1}{12!}.$$

Это число так мало, что не может повлиять на девятую цифру, и потому, допуская возможную ошибку в последней цифре вышеприведенного значения, мы получаем для e следующее приближенное равенство с восемью верными цифрами:

$$e \approx 2{,}7182818.$$

Число e иррационально. Чтобы это доказать, предположим противное: допуская, что $e = \frac{p}{q}$, где p и q — целые числа, и затем, приходя к противоречию, мы должны будем заключить о нелепости сделанного предположения. Поскольку мы знаем, что $2 < e < 3$, e не может быть целым числом, а потому q по меньшей мере должно быть равно 2. Умножим обе части тождества (6)

на $q! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot q$; получим

$$e \cdot q! = p \cdot 2 \cdot 3 \cdot \ldots \cdot (q-1) = \\ = [q! + q! + 3 \cdot 4 \cdot \ldots \cdot q + 4 \cdot 5 \cdot \ldots \cdot q + \ldots + (q-1)q + q + 1] + \\ + \frac{1}{q+1} + \frac{1}{(q+1)(q+2)} + \ldots \quad (7)$$

В левой части мы, очевидно, имеем целое число. В правой части слагаемое в квадратных скобках также есть целое число. Остаток же в правой части есть положительное число, меньшее $\frac{1}{2}$, и значит, не есть целое число. В самом деле, $q \geqslant 2$, а следовательно, члены ряда $\frac{1}{q+1} + \ldots$ не превышают соответственно членов геометрической прогрессии $\frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \ldots$, сумма которых равна $\frac{1}{3} \cdot \frac{1}{1 - \frac{1}{3}} = \frac{1}{2}$. Таким образом, формула (7) противоречива: целое число в левой части не может быть равно числу в правой части, так как это последнее, являясь суммой целого числа и положительного числа, меньшего $\frac{1}{2}$, не есть целое число.

4. Число π. Как известно из школьной математики, длина окружности, радиус которой равен единице, может быть определена как предел последовательности длин периметров правильных многоугольников при бесконечном увеличении числа их сторон. Определенная таким образом длина окружности обозначается символом 2π. Точнее, если через p_n обозначить длину вписанного, а через q_n длину описанного правильного n-угольника, то имеют место неравенства

$$p_n < 2\pi < q_n.$$

Более того, когда n возрастает, обе последовательности p_n и q_n приближаются монотонно к 2π, и на каждом этапе мы получаем все меньшую ошибку в том приближении, которое p_n и q_n дают для числа 2π.

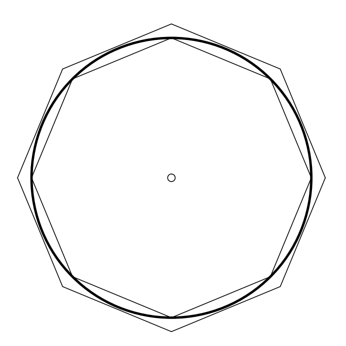


Рис. 167. Приближение окружности многоугольниками

На стр. 145 мы получили выражение

$$p_{2^m} = 2^m \sqrt{2 - \sqrt{2 + \sqrt{2 + \ldots}}},$$

содержащее $m - 1$ перекрывающихся квадратных корней. Эту формулу можно использовать для подсчета приближенного значения числа 2π.

Упражнения. 1) Найдите приближенное значение π, даваемое числами p_4, p_8 и p_{16}.

*2) Найдите формулу для q_{2^m}.

*3) С помощью этой формулы вычислите q_4, q_8 и q_{16}. Зная величины p_{16} и q_{16}, установите границы, между которыми должно лежать число π.

Что же это за число π? Неравенство $p_n < 2\pi < q_n$ дает на это полный ответ при развертывании последовательности вложенных интервалов, которые стягиваются к точке 2π. И все же этот ответ оставляет желать еще чего-то, поскольку он ничего не говорит о природе π как действительного числа: является ли оно рациональным или иррациональным, алгебраическим или трансцендентным? Как мы уже указывали на стр. 127, число π есть число трансцендентное, а следовательно, и иррациональное. В противоположность доказательству для e доказательство иррациональности π, впервые данное Ламбертом (1728–1777), в достаточной мере трудно и здесь приведено не будет. Однако ряд сведений о числе π мы можем сообщить. Имея в виду, что целые числа являются существенной основой математики, мы можем задать вопрос: связывается ли число π сколько-нибудь просто и непосредственно с целыми числами? Десятичное разложение числа π, хотя и вычисленное с несколькими сотнями знаков, не обнаруживает ни малейшей закономерности. Это и не удивительно: ведь π и число 10 не имеют между собой ничего общего. Однако Эйлер (XVIII в.) и другие нашли изящные выражения, связывающие число π с целыми числами с помощью бесконечных рядов и произведений. Простейшей из таких формул является, вероятно, следующая:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots,$$

выражающая $\frac{\pi}{4}$ как предел при возрастающем n сумм

$$s_n = 1 - \frac{1}{3} + \frac{1}{5} - \ldots + (-1)^n \frac{1}{2n+1}.$$

Эту формулу мы выведем в главе VIII. Вот другой бесконечный ряд, который может служить для вычисления π:

$$\frac{\pi^2}{6} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \ldots$$

И еще одно удивительное выражение для π было открыто английским математиком Джоном Уоллисом (1616–1703). Его формула утверждает следующее:

$$\left\{\frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \ldots \cdot \frac{2n}{2n-1} \cdot \frac{2n}{2n+1}\right\} \to \frac{\pi}{2} \quad \text{при} \quad n \to \infty.$$

В сокращенном виде она часто записывается так:

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \frac{8}{7} \cdot \frac{8}{9} \ldots$$

(выражения, подобные стоящему в правой части, называются *бесконечными произведениями*).

Доказательство последних двух формул можно найти в любом достаточно полном курсе анализа.

***5. Непрерывные дроби.** Интересные бесконечные процессы возникают в связи с непрерывными дробями. Конечная непрерывная дробь, такая как

$$\frac{57}{17} = 3 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{5}}},$$

представляет собой некоторое рациональное число. На стр. 67 мы показали, что каждое рациональное число может быть выражено в такой форме с помощью алгоритма Евклида. Однако в случае иррациональных чисел алгоритм не заканчивается после конечного числа операций. Напротив, он ведет к последовательности все более «длинных» дробей, из которых каждая представляет собой рациональное число. В частности, все действительные алгебраические числа (см. стр. 123) степени 2 могут быть выражены таким образом. Рассмотрим, например, число $x = \sqrt{2} - 1$, являющееся корнем квадратного уравнения

$$x^2 + 2x = 1, \qquad \text{или} \qquad x = \frac{1}{2+x}.$$

Если в правой части заменить x снова дробью $x = \dfrac{1}{2+x}$, то это дает выражение

$$x = \cfrac{1}{2 + \cfrac{1}{2+x}},$$

а затем

$$x = \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2+x}}}$$

и т. д., так что после n «шагов» получим равенство

$$\left. x = \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \dots \; \dots + \cfrac{1}{2+x}}}} \right\} \quad (n \text{ «шагов»}).$$

Если n стремится к бесконечности, мы получим «бесконечную непрерывную дробь»

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \dots}}}.$$

Эта замечательная формула связывает число $\sqrt{2}$ с целыми числами гораздо более удивительным образом, чем это делает десятичное разложение $\sqrt{2}$, которое не обнаруживает никакой правильности в чередовании десятичных знаков.

Для положительного корня любого квадратного уравнения вида

$$x^2 = ax + 1, \qquad \text{или} \qquad x = a + \frac{1}{x},$$

мы получаем разложение

$$x = a + \cfrac{1}{a + \cfrac{1}{a + \cfrac{1}{a + \ldots}}}.$$

Например, полагая $a = 1$, мы находим

$$x = \frac{1}{2}(1 + \sqrt{5}) = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \ldots}}}.$$

Эти примеры являются частными случаями общей теоремы, утверждающей, что *действительные корни квадратного уравнения с целыми коэффициентами разлагаются в периодическую непрерывную дробь*, подобно тому как рациональные числа разлагаются в периодические десятичные дроби.

Эйлер сумел найти почти столь же простые разложения в непрерывные дроби для чисел e и π. Приведем их без доказательств:

$$e = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{4 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{6 + \ldots}}}}}}}};$$

$$e = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{2}{3 + \cfrac{3}{4 + \cfrac{4}{5 + \ldots}}}}}; \qquad \frac{\pi}{4} = \cfrac{1}{1 + \cfrac{1^2}{2 + \cfrac{3^2}{2 + \cfrac{5^2}{2 + \cfrac{7^2}{2 + \cfrac{9^2}{2 + \ldots}}}}}}.$$

§ 3. Пределы при непрерывном приближении

1. Введение. Общие определения. В § 2, пункт 1, нам удалось дать точное определение утверждению: «Последовательность a_n (т. е. функция $a_n = F(n)$ натурального переменного n) имеет предел a при n, стремящемся к бесконечности». Теперь мы дадим соответствующее определение утверждению: «Функция $u = f(x)$ непрерывной переменной x имеет предел a при стремлении x к значению x_1».

В интуитивной форме понятие предела при непрерывном приближении независимого переменного x употреблялось уже в § 1, пункт 5, когда

нужно было установить, непрерывна ли рассматриваемая функция в данной точке.

Начнем опять с частного примера. Функция $f(x) = \dfrac{x+x^3}{x}$ определена для всех значений x, не равных нулю; при этом последнем значении x знаменатель уничтожается. Если мы вычертим график функции $y = f(x)$ для значений x в окрестности точки 0, то станет очевидным, что при x, «стремящемся» к 0 с любой стороны, соответствующие значения $u = f(x)$ «стремятся» к пределу 1. Для того чтобы дать точное описание этого факта, найдем явную формулу разности между значением функции $f(x)$ и постоянного числа 1:

$$f(x) - 1 = \frac{x+x^3}{x} - 1 = \frac{x+x^3-x}{x} = \frac{x^3}{x}.$$

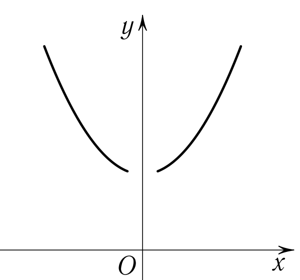


Рис. 168. $u = \dfrac{x+x^3}{x}$

Если мы условимся рассматривать лишь значения x, близкие к 0, но не равные самому нулю (для которого функция $f(x)$ даже не определена), мы можем разделить числитель и знаменатель на x и получить более простую формулу

$$f(x) - 1 = x^2.$$

Ясно, что эту разность мы можем сделать *сколь угодно малой*, ограничивая изменение переменной x *достаточно малой* окрестностью значения $x = 0$. Так, например, при $x = \pm\dfrac{1}{10}$ имеем $f(x) - 1 = \dfrac{1}{100}$; при $x = \pm\dfrac{1}{100}$ имеем $f(x) - 1 = \dfrac{1}{10\,000}$, и т. д. Вообще, если ε есть некоторое положительное число, то, как бы мало оно ни было, разность между $f(x)$ и 1 будет меньше чем ε, если только расстояние точки x от точки 0 меньше числа $\delta = \sqrt{\varepsilon}$.

В самом деле, если $|x| < \sqrt{\varepsilon}$, то

$$|f(x) - 1| = |x^2| < \varepsilon.$$

Аналогия с нашим определением предела последовательности полная. На стр. 313 мы дали определение: *последовательность a_n имеет предел a при n, стремящемся к бесконечности, если каждому положительному числу ε, как бы мало оно ни было, можно поставить в соответствие такое целое N (зависящее от ε), что неравенство*

$$|a_n - a| < \varepsilon$$

выполняется для всех n, удовлетворяющих неравенству

$$n \geqslant N.$$

В случае функции $f(x)$ непрерывного переменного x при x, стремящемся к некоторому конечному значению x_1, мы просто слова «достаточно

большое n» (что характеризуется числом N) заменяем словами «достаточно близко к x_1» (что характеризуется числом δ) и приходим к следующему определению предела при непрерывном приближении, впервые данному Коши около 1820 г.: *функция $f(x)$ имеет предел a, когда x стремится к значению x_1, если каждому положительному числу ε, как бы мало оно ни было, можно поставить в соответствие такое положительное число δ (зависящее от ε), что*

$$|f(x) - a| < \varepsilon$$

для всех значений $x \neq x_1$, удовлетворяющих неравенству

$$|x - x_1| < \delta.$$

Если это имеет место, принято писать

$$f(x) \to a \quad \text{при} \quad x \to x_1.$$

В случае функции $f(x) = \dfrac{x + x^3}{x}$ мы выше показали, что эта функция $f(x)$ имеет предел 1 при x, стремящемся к значению $x_1 = 0$. В этом случае достаточно было всегда выбирать $\delta = \sqrt{\varepsilon}$.

2. Замечания по поводу понятия предела. ε-δ-определение предела — результат столетних попыток и блужданий; оно кратко воплощает результат неустанных усилий поставить понятие предела на здоровую математическую основу. Важнейшие понятия анализа — производная и интеграл — могут быть определены не иначе, как с помощью перехода к пределу. Но ясное понимание и строгое определение самого понятия предела долгое время казались непреодолимо трудными.

При изучении движения в частности и какого бы то ни было изменения в общем случае математики XVII и XVIII столетий принимали, как нечто достаточно наглядное и не подлежащее дальнейшему анализу, концепцию величины x, меняющейся и в своем непрерывном течении приближающейся к предельному значению x_1. Они рассматривали другую величину $u = f(x)$, зависящую от времени или от какой-нибудь другой зависящей от времени величины. Оставалось все же проблемой: какой точный математический смысл следует приписывать представлению о том, что $f(x)$ «стремится» или «приближается» к определенному значению a, когда x движется к x_1?

Однако еще со времен Зенона и его парадоксов все попытки дать точную математическую формулировку интуитивному физическому или метафизическому понятию непрерывного движения были безуспешными. Нет затруднений в продвижении шаг за шагом по дискретной последовательности значений a_1, a_2, a_3, ... Но когда приходится иметь дело с непрерывной переменной x, пробегающей целый интервал значений на числовой оси, то описание того, как x «приближается» к заданному

значению x_1, затруднено тем, что принимаемые значения из интервала не могут быть указаны последовательно в порядке их возрастания. В самом деле, точки прямой представляют везде плотное множество, и не существует точки, «следующей» за данной. Остается неизбежное расхождение между интуитивной идеей и точным математическим языком, предназначенным для того, чтобы описывать ее основные линии в научных, логических терминах. Парадоксы Зенона ярко обнаруживают это несоответствие.

Существенным достижением Коши является то, что он ясно осознал, что, поскольку дело касается математических понятий, всякая ссылка на интуитивное представление о непрерывном движении должна быть отброшена. Как случается нередко, подлинный научный прогресс был осуществлен тогда, когда последовал отказ от попыток прибегать к метафизическим объяснениям и было принято решение вести рассуждение, оставаясь на почве строго математических понятий, соответствующих «наблюдаемым фактам» в физике. Если мы проанализируем логически, что надлежит понимать под «непрерывным приближением» и какие существуют способы для того, чтобы в каждом отдельном случае проверить, имеет ли место таковое, то мы вынуждены будем принять именно то самое определение, которое дано Коши, и никакое иное. Это определение — статическое; оно не опирается на интуитивную идею движения. Более того, только такое статическое определение позволяет подвергнуть точному математическому анализу само непрерывное движение и разрешает парадоксы Зенона, по крайней мере в той их части, которая относится к математике.

В определении с помощью ε, δ независимое переменное не «движется»; оно не «стремится» и не «приближается» к пределу x_1 в каком бы то ни было физическом смысле. Правда, эти обороты речи, как и символ $\to$, сохраняются, причем математик вовсе не обязан отказываться от тех, в общем-то весьма полезных, интуитивных представлений, которые с ними связываются. Но когда в частном случае нужно дать ответ на вопрос, существует предел или не существует, то приходится прибегнуть именно к определению с помощью ε, δ. Спрашивать о том, насколько удовлетворительно это определение соответствует интуитивному «динамическому» представлению о стремлении к пределу, можно с таким же правом, как и о том, насколько удовлетворительно аксиомы геометрии описывают то, что мы называем пространством (в интуитивном смысле).

Обе формулировки в какой-то степени предоставляют возможность работать воображению, и вместе с тем обе они создают адекватную математическую основу для дальнейшего логического построения.

Как и в случае предела последовательности, ключ к правильному пониманию определения Коши лежит в обращении «естественного» порядка, в котором рассматриваются переменные. Прежде мы отмечаем

границу ε для зависимого переменного, а уже потом стремимся определить подходящую границу δ для независимого переменного. Когда мы говорим, что «$f(x) \to a$ при $x \to x_1$», то лишь сокращенно высказываем ту мысль, что этот процесс может быть выполнен для любого положительного числа ε. В частности, ни одна из частей этого утверждения (например, «$x \to x_1$») не имеет смысла сама по себе.

Еще нужно подчеркнуть следующее. Заставляя x «стремиться» к x_1, мы можем позволить x быть больше или меньше, чем x_1, но возможность равенства явно исключается требованием $x \neq x_1$: x стремится к x_1, но никогда не принимает значения x_1. Таким образом, мы можем применять наше определение к функциям, не определенным вовсе при $x = x_1$, но имеющим тот или иной предел при x, стремящемся к x_1, например, к функции $f(x) = \frac{x + x^3}{x}$, рассмотренной на стр. 325. Исключение значения $x = x_1$ как раз соответствует тому факту, что, рассматривая последовательности a_n при $n \to \infty$ $\left(\text{например, предел } a_n = \frac{1}{n}\right)$, мы никогда не подставляем в формулу значения $n = \infty$.

Однако, что касается функции $f(x)$, то когда x стремится к x_1, ей не запрещено стремиться к пределу a таким образом, что при некоторых значениях $x \neq x_1$ осуществляется равенство $f(x) = a$. Например, рассматривая функцию $f(x) = \frac{x}{x}$ при x, стремящемся к 0, мы никогда не позволяем x быть равным 0, но зато, напротив, равенство $f(x) = 1$ справедливо при всех $x \neq 0$, и предел a существует и равен 1 в точном согласии с определением.

3. Предел $\frac{\sin x}{x}$. Если x обозначает угол в радианном измерении, то выражение $\frac{\sin x}{x}$ определено для всех значений x, за исключением значения $x = 0$, при котором оно принимает вид не имеющего смысла символа $\frac{0}{0}$. С помощью таблиц тригонометрических функций читатель может подсчитать значение частного $\frac{\sin x}{x}$ для малых значений x. Эти таблицы обычно даются для градусного измерения углов; мы напоминаем (см. § 1, пункт 2), что градусная мера x связана с радианной мерой y следующим соотношением: $x = \frac{\pi}{180} y = 0{,}01745y$ (с точностью до пятого десятичного знака). Из четырехзначных таблиц мы находим следующие

значения:

	x	$\sin x$	$\dfrac{\sin x}{x}$
10°	0,1745	0,1736	0,9948
5°	0,0873	0,0872	0,9988
2°	0,0349	0,0349	1,0000
1°	0,0175	0,0175	1,0000

Хотя точность чисел здесь ограничивается четырьмя знаками, все же эти данные приводят к мысли, что

$$\frac{\sin x}{x} \to 1 \quad \text{при} \quad x \to 0. \tag{1}$$

Сейчас мы дадим строгое доказательство этому предельному соотношению.

В силу определения тригонометрических функций с помощью единичного круга, мы имеем следующие соотношения для величины x, являющейся радианной мерой угла BOC (см. рис. 169) при ограничении $0 < x < \frac{\pi}{2}$.

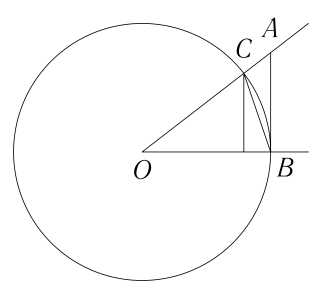


Рис. 169. Основное тригонометрическое неравенство

Площадь треугольника $OBC = \frac{1}{2} \cdot 1 \cdot \sin x$.

Площадь кругового сектора $OBC = \frac{1}{2} \cdot x$.

Площадь треугольника $OBA = \frac{1}{2} \cdot 1 \cdot \operatorname{tg} x$.

Отсюда вытекает двойное неравенство

$$\sin x < x < \operatorname{tg} x.$$

Деля на $\sin x$, получим, далее,

$$1 < \frac{x}{\sin x} < \frac{1}{\cos x},$$

или

$$\cos x < \frac{\sin x}{x} < 1. \tag{2}$$

Но, с другой стороны,

$$1 - \cos x = (1 - \cos x) \cdot \frac{1 + \cos x}{1 + \cos x} = \frac{1 - \cos^2 x}{1 + \cos x} = \frac{\sin^2 x}{1 + \cos x} < \sin^2 x.$$

Так как $\sin x < x$, то отсюда следует, что

$$1 - \cos x < x^2, \tag{3}$$

или

$$1 - x^2 < \cos x.$$

Совместно с неравенством (2) это дает окончательно нужные нам неравенства

$$1 - x^2 < \frac{\sin x}{x} < 1. \tag{4}$$

Мы предполагаем, что $0 < x < \frac{\pi}{2}$; однако неравенства (4) справедливы и при условии $-\frac{\pi}{2} < x < 0$, поскольку $\frac{\sin(-x)}{(-x)} = \frac{-\sin x}{-x} = \frac{\sin x}{x}$ и $(-x)^2 = x^2$.

Предельное соотношение (1) вытекает немедленно из неравенств (4). В самом деле, разность между $\frac{\sin x}{x}$ и 1 меньше, чем x^2, а x^2 может быть сделано меньше, чем любое число ε, если только взять $|x| < \delta = \sqrt{\varepsilon}$.

Упражнения. 1) Выведите из неравенства (3) предельное соотношение

$$\frac{1 - \cos x}{x} \to 0 \quad \text{при} \quad x \to 0.$$

Найдите пределы при $x \to 0$ следующих функций:

2) $\dfrac{\sin^2 x}{x}$, 3) $\dfrac{\sin x}{x(x-1)}$, 4) $\dfrac{\operatorname{tg} x}{x}$, 5) $\dfrac{\sin ax}{x}$, 6) $\dfrac{\sin ax}{\sin bx}$, 7) $\dfrac{x \sin x}{1-\cos x}$, 8) $\dfrac{\sin x}{x}$, предполагая, что x измеряется в градусах,

$$9)\ \frac{1}{x} - \frac{1}{\operatorname{tg} x}, \qquad 10)\ \frac{1}{\sin x} - \frac{1}{\operatorname{tg} x}.$$

4. Пределы при $x \to \infty$. Если непрерывная переменная x достаточно велика, то функция $f(x) = \dfrac{1}{x}$ становится произвольно малой, или «стремится к 0». В самом деле, поведение этой функции при возрастающем x по существу то же самое, что и поведение последовательности $\dfrac{1}{n}$ при возрастании n. Мы вводим общее определение: *функция $f(x)$ имеет предел a при x, стремящемся к бесконечности*, и записываем это в форме

$$f(x) \to a \quad \text{при} \quad x \to \infty,$$

если, как бы мало ни было положительное число ε, можно к нему подобрать такое положительное число K (зависящее от ε), что неравенство

$$|f(x) - a| < \varepsilon$$

выполняется при условии $|x| > K$ (сравните с соответствующим определением на стр. 313).

В случае функции $f(x) = \dfrac{1}{x}$, для которой $a = 0$, достаточно выбрать $K = \dfrac{1}{\varepsilon}$, в чем читатель может убедиться немедленно.

Упражнения. 1) Покажите, что с точки зрения вышеприведенного определения, утверждение

$$f(x) \to a \quad \text{при} \quad x \to \infty$$

эквивалентно следующему:

$$f(x) \to a \quad \text{при} \quad \frac{1}{x} \to 0.$$

Докажите, что имеют место следующие предельные соотношения при $x \to \infty$:

2) $\dfrac{x+1}{x-1} \to 1$ при $x \to \infty$, 3) $\dfrac{x^2+x+1}{x^2-x-1} \to 1$ при $x \to \infty$,

4) $\dfrac{\sin x}{x} \to 0$ при $x \to \infty$, 5) $\dfrac{x+1}{x^2+1} \to 0$ при $x \to \infty$,

6) $\dfrac{\sin x}{x + \cos x} \to 0$ при $x \to \infty$, 7) $\dfrac{\sin x}{\cos x}$ не имеет предела при $x \to \infty$.

8) Дайте определение «$f(x) \to \infty$ при $x \to \infty$». Приведите пример.

Имеется следующая разница между случаем функции $f(x)$ и случаем последовательности a_n. В случае последовательности n может стремиться к бесконечности не иначе, как возрастая, тогда как в случае функции переменная x, неограниченно возрастая, имеет право принимать как положительные, так и отрицательные значения. Если желательно направить внимание на

поведение функции $f(x)$ только при больших *положительных* значениях, то условие $|x| > K$ мы должны заменить условием $x > K$; напротив, для случая больших по абсолютной величине отрицательных значений x вводим условие $x < -K$. Чтобы символизировать эти два способа «одностороннего» стремления к бесконечности, мы пишем, соответственно,

$$x \to +\infty, \qquad x \to -\infty.$$

§ 4. Точное определение непрерывности

В § 1, пункт 5, мы ввели следующее определение непрерывности функции: функция $f(x)$ непрерывна в точке $x = x_1$, если при стремлении x к x_1 величина $f(x)$ стремится к пределу, равному $f(x_1)$. Если мы проанализируем эту формулировку, то увидим, что она подразумевает выполнение следующих двух требований:

а) существует предел a функции $f(x)$ при стремлении переменной x к x_1,

б) этот предел a должен быть равен $f(x_1)$.

Если в определении предела на стр. 326 мы подставим вместо a его значение $f(x_1)$, то условие непрерывности принимает следующий вид: *функция $f(x)$ непрерывна при $x = x_1$, если, как бы мало ни было положительное число ε, можно подобрать такое положительное число δ (зависящее от ε), что неравенство*

$$|f(x) - f(x_1)| < \varepsilon$$

будет выполнено для всех x, удовлетворяющих условию

$$|x - x_1| < \delta$$

(ограничение $x \neq x_1$, введенное в определении предела, здесь излишне, поскольку неравенство $|f(x) - f(x_1)| < \varepsilon$ при $x = x_1$ удовлетворяется автоматически).

В качестве примера постараемся установить непрерывность функции $f(x) = x^3$, скажем, в точке $x_1 = 0$. Мы имеем

$$f(x_1) = 0^3 = 0.$$

Выберем теперь маленькое положительное число ε, например, $\varepsilon = \frac{1}{1000}$. Мы должны показать, что, ограничивая значения x числами, достаточно близкими к 0, получим соответствующие значения функции $f(x)$, отличающиеся от 0 меньше, чем на $\frac{1}{1000}$, т. е. заключенные между $-\frac{1}{1000}$ и $+\frac{1}{1000}$. Мы сразу видим, что значения $f(x)$ не выйдут из этих границ, если мы ограничим изменение x значениями, отличающимися от 0 меньше чем на $\delta = \sqrt[3]{\frac{1}{1000}} = \frac{1}{10}$; в самом деле, если $|x| < \frac{1}{10}$, то $|f(x)| = |x^3| <$

$\frac{1}{1000}$. Совершенно так же мы можем взять вместо $\varepsilon = \frac{1}{1000}$ любое меньшее значение $\varepsilon = 10^{-4}$, 10^{-5} и т. д.; числа $\delta = \sqrt[3]{\varepsilon}$ будут удовлетворять нашему требованию, так как из неравенства $|x| < \sqrt[3]{\varepsilon}$ следует неравенство $|f(x)| = |x^3| < \varepsilon$.

Основываясь на определении непрерывности с помощью ε, δ, можно доказать аналогично, что все полиномы, рациональные функции и тригонометрические функции непрерывны в любой точке, за исключением, может быть, тех изолированных значений x, около которых функции становятся бесконечными.

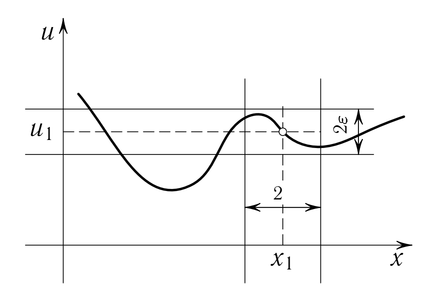


Рис. 170. Функция, непрерывная в точке $x = x_1$

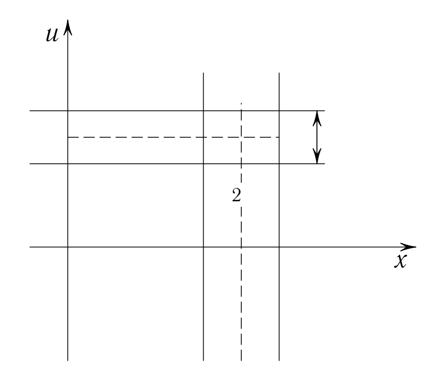


Рис. 171. Функция имеет разрыв в точке $x = x_1$

Связывая определение непрерывности с графиком функции $u = f(x)$, можно придать ему следующую геометрическую форму. Выберем некоторое положительное число ε и начертим прямые, параллельные оси x на высоте $f(x_1) - \varepsilon$ и $f(x_1) + \varepsilon$ над ней. Тогда должно найтись такое положительное число δ, что вся часть графика, лежащая внутри вертикальной полоски шириной в 2δ около x_1, содержится также и в горизонтальной полоске шириной в 2ε около $f(x_1)$. Рис. 170 показывает функцию, непрерывную в точке x_1, в то время как рис. 171 показывает функцию, имеющую разрыв в этой точке. В последнем случае, как бы ни была узка вертикальная полоска около x_1, она всегда будет содержать часть графика, лежащую вне горизонтальной полоски ширины 2ε.

Если я утверждаю, что данная функция $u = f(x)$ непрерывна в точке $x = x_1$, то это значит, что я беру на себя по отношению к вам следующие обязательства: вы можете выбрать любое положительное число ε, сколь угодно малое, но определенное. Тогда я обязуюсь подыскать такое положительное число δ, чтобы неравенство $|x - x_1| < \delta$ влекло за собой неравенство $|f(x) - f(x_1)| < \varepsilon$. Но при этом я не обязуюсь найти такое число δ, которое подошло бы ко всякому ε, которое вы назовете *потом: мой* выбор δ зависит от *вашего* выбора ε. Если вы можете выбрать хоть одно ε, для которого я не смогу подобрать подходящего δ, то моя игра проиграна — мое утверждение опро-

вергнуто. Для того чтобы доказать, что я могу выполнить мое обязательство в конкретном случае некоторой функции $u = f(x)$, мне нужно построить явно такую положительную функцию

$$\delta = \varphi(\varepsilon),$$

определенную для всякого положительного числа ε, для которой я могу доказать, что из неравенства $|x - x_1| < \delta$ всегда следует неравенство $|f(x) - f(x_1)| < \varepsilon$. В случае функции $u = f(x) = x^3$ при $x_1 = 0$ функцией $\delta = \varphi(\varepsilon)$ была $\delta = \sqrt[3]{\varepsilon}$.

Упражнения. 1) Докажите, что $\sin x$ и $\cos x$ — непрерывные функции.
2) Докажите непрерывность функций $\dfrac{1}{1+x^2}$ и $\sqrt{1+x^2}$.

Теперь становится ясным, что определение непрерывности с помощью ε, δ не находится в противоречии с тем, что мы могли бы назвать «наблюдаемыми фактами», относящимися к функциям. Таким образом, оно соответствует основному принципу современной науки, который выдвигает в качестве критерия полезности некоторого понятия или «существования» явления (в научном смысле) возможность непосредственно его наблюдать (по крайней мере в принципе) или сводить его к фактам, доступным наблюдению.

§ 5. Две основные теоремы о непрерывных функциях

1. Теорема Больцано. Бернард Больцано (1781–1848), католический священник, знаток схоластической философии, был одним из первых, кто ввел в математический анализ современное понятие строгости. Его замечательная книжка «Paradoxien des Unendlichen» появилась в 1850 г. Здесь впервые было признано, что многие казалось бы очевидные утверждения, касающиеся непрерывных функций, могут и должны быть доказаны, если имеется в виду применять их во всей их общности. Примером этого может служить следующая теорема о функциях одного переменного.

Непрерывная функция переменного x, положительная при некотором значении x и отрицательная при некотором другом значении x из замкнутого интервала непрерывности $a \leqslant x \leqslant b$, должна обращаться в нуль при некотором промежуточном значении x. Итак, если функция $f(x)$ непрерывна при изменении x от a до b, и при этом $f(a) < 0$ и $f(b) > 0$, то существует такое значение α переменного x, что $a < \alpha < b$ и $f(\alpha) = 0$.

Теорема Больцано прекрасно согласуется с нашим интуитивным представлением о непрерывной кривой, которая неизбежно должна пересечь в какой-нибудь точке ось x, чтобы перейти с одной ее стороны

на другую. Что, напротив, это не обязательно в случае разрывной функции, показывает рис. 157 на стр. 305.

***2. Доказательство теоремы Больцано.** Дадим строгое доказательство этой теоремы. Если следовать Гауссу и другим великим математикам, то можно принять этот факт и без доказательства. Нашей целью является сведение этой теоремы к основным свойствам системы действительных чисел, в частности, к постулату Дедекинда—Кантора о стягивающихся отрезках (стр. 88). Для этого рассмотрим отрезок I, $a \leqslant x \leqslant b$, в котором задана функция $f(x)$, и разобьем его на два средней точкой $x_1 = \dfrac{a+b}{2}$. Если в этой средней точке мы получим $f(x_1) = 0$, то доказывать больше уже нечего. Если, однако, $f(x_1) \neq 0$, то $f(x_1)$ должно быть или больше, или меньше нуля. В обоих случаях одна из половинок отрезка I будет снова обладать тем свойством, что знаки значений функции $f(x)$ на его концах различны.

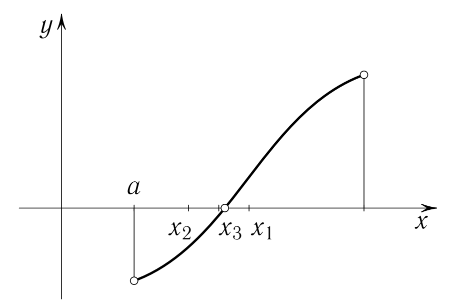


Рис. 172. Теорема Больцано

Обозначим этот отрезок через I_1. Мы повторим этот процесс, деля отрезок I_1 пополам; тогда либо мы в середине I_1 имеем $f(x) = 0$, либо мы можем выбрать такую половину I_2 отрезка I_1, для которой опять знаки значений функции на двух концах различны. Повторяя эту процедуру, мы, в конце концов, или найдем после конечного числа делений точку, в которой $f(x) = 0$, или получим последовательность стягивающихся отрезков I_1, I_2, I_3, ... В последнем случае постулат Дедекинда—Кантора обеспечивает существование в исходном отрезке I такой точки α, которая принадлежит всем отрезкам сразу. Мы утверждаем, что $f(\alpha) = 0$, так что α и будет той точкой, существование которой нужно доказать.

До сих пор предположение о непрерывности функции $f(x)$ использовано еще не было. Сейчас придется на него сослаться, заканчивая доказательство способом от противного. Мы докажем, что $f(\alpha) = 0$, допуская противоположное и приходя затем к противоречию. Предположим, что $f(\alpha) \neq 0$: пусть, например, $f(\alpha) = 2\varepsilon > 0$. Так как функция $f(x)$ непрерывна, то мы найдем (может быть, маленький) отрезок J длины 2δ с центром в точке α — такой, что значение функции $f(x)$ во всем промежутке J отличается от $f(\alpha)$ меньше чем на ε. Затем, так как $f(\alpha) = 2\varepsilon$, то мы можем быть уверены, что $f(x) > \varepsilon$ в каждой точке J, т. е. что $f(x) > 0$ в отрезке J. Но отрезок J фиксирован, и если n достаточно велико, то маленький отрезок I_n должен непременно попасть внутрь J, поскольку последовательность длин I_n стремится к нулю. В этом заключается противоречие: в самом деле, из того, каким образом был выбран промежуток I_n, вытекает, что функция $f(x)$ имеет противоположные знаки в двух конечных точках каждого промежутка I_n, так что функция $f(x)$

принимает отрицательные значения где-то в промежутке J. Отсюда следует нелепость предположения $f(\alpha) > 0$, а также (совершенно таким же образом) $f(\alpha) < 0$; следовательно, доказано, что $f(\alpha) = 0$.

3. Теорема Вейерштрасса об экстремальных значениях. Другой существенный и интуитивно ясный факт, касающийся непрерывных функций, был сформулирован Карлом Вейерштрассом (1815–1897), который, возможно, более чем кто-либо другой является ответственным за современное стремление к строгости в математическом анализе. Эта теорема утверждает: *если функция $f(x)$ непрерывна в интервале I, $a \leqslant x \leqslant b$, не исключая также и конечных точек интервала a и b, то в интервале I должна существовать по крайней мере одна точка, в которой функция $f(x)$ достигает своего наибольшего значения M, и другая точка, в которой функция $f(x)$ достигает своего наименьшего значения m.* Говоря интуитивно, это значит, что график непрерывной функции $u = f(x)$ должен иметь по крайней мере одну наивысшую и одну наинизшую точки.

Существенно отметить, что это утверждение может быть неверным, если функция $u = f(x)$ перестает быть непрерывной в конечных точках промежутка I. Например, функция $f(x) = \dfrac{1}{x}$ не имеет наибольшего значения в промежутке $0 < x \leqslant 1$, хотя она и непрерывна внутри промежутка. И вместе с тем разрывная функция вовсе не обязательно достигает наибольшего и наименьшего значений, даже если она ограниченная. Рассмотрим, например, «чрезвычайно» разрывную функцию $f(x)$, определенную следующим образом:

$$f(x) = x \qquad \text{при иррациональном } x,$$
$$f(x) = \frac{1}{2} \qquad \text{при рациональном } x$$

в промежутке $0 \leqslant x \leqslant 1$. Все значения, которые принимает эта функция, заключены между 0 и 1. Среди них имеются сколь угодно близкие к 0 и 1: они получаются, если x будем выбирать иррациональным и достаточно близким к 0 или 1. Но $f(x)$ никогда не может быть *равным* ни 0, ни 1, поскольку для рациональных x мы имеем $f(x) = \dfrac{1}{2}$, а для иррациональных мы имеем $f(x) = x$. Итак, значения 0 и 1 ни в какой точке не достигаются.

* Теорема Вейерштрасса может быть доказана почти таким же образом, как и теорема Больцано. Разобьем интервал I на два замкнутых полуинтервала I' и I'' и фиксируем наше внимание на I', как на интервале, в котором следует искать наибольшее значение функции $f(x)$, *если только в интервале I'' не найдется такой точки α, что $f(\alpha)$ больше всех значений функции $f(x)$ в интервале I'*; *в этом последнем случае мы выберем интервал I''*. Тот интервал, который мы выбрали, обозначим через I_1. Поступим теперь с

интервалом I_1 точно так же, как мы поступали с I; пусть при этом получим интервал I_2, и т. д. Этот процесс определит последовательность I_1, I_2, I_3, ..., I_N, ... вложенных интервалов, которые все содержат некоторую точку z. Мы докажем, что значение функции в этой точке, $f(z) = M$, есть наибольшее из всех значений функции $f(x)$, достигаемых в исходном интервале, т. е. что не может существовать такой точки s, что $f(s) > M$. Предположим, что нашлась бы точка s, удовлетворяющая условию $f(s) = M + 2\varepsilon$, где ε есть некоторое (может быть, и очень маленькое) положительное число. В силу непрерывности функции $f(x)$ мы можем точку z окружить маленьким интервалом K, не захватывающим точки s, и притом таким, что в интервале K значения функции $f(x)$ отличаются от $f(z) = M$ меньше чем на ε, так что в нем мы непременно будем иметь $f(x) < M + \varepsilon$. Но при достаточно больших n интервал I_n лежит внутри интервала K, а вместе с тем интервал I_n был определен так, что ни одно значение $f(x)$ при x, лежащем вне интервала I_n, не может превзойти значений функции $f(x)$ в точках x из этого интервала.

Но точка s лежит вне интервала I_n и в ней $f(s) > M + \varepsilon$, тогда как в интервале K, а тем самым и в интервале I_n, мы имеем $f(x) < M + \varepsilon$. Таким образом, мы пришли к противоречию.

Существование по крайней мере одного наименьшего значения m может быть доказано тем же самым методом; впрочем, оно является следствием предыдущего, так как наименьшее значение функции $f(x)$ является наибольшим значением функции $g(x) = -f(x)$.

Теорема Вейерштрасса может быть доказана аналогичным образом и для непрерывных функций от двух или большего числа переменных x, y, ... В этом случае придется вместо замкнутых интервалов (со включением конечных точек) брать замкнутые области, например, прямоугольники в плоскости x, y (со включением контура).

Упражнение. В каком пункте доказательств теорем Больцано и Вейерштрасса мы воспользовались предположением, что функция $f(x)$ определена и непрерывна во всем отрезке (*замкнутом*) $a \leqslant x \leqslant b$, а не только при $a < x \leqslant b$ или $a < x < b$?

Доказательства теорем Больцано в Вейерштрасса носят явно неконструктивный характер. Они не предоставляют метода для «эффективного» нахождения положения нулевой точки или наибольшего и наименьшего значения функции с заранее назначенной степенью точности в результате конечного числа операций. Доказано только лишь само существование, или, вернее, абсурдность несуществования, упомянутых значений. Это обстоятельство представляет собой еще один важный пункт, против которого «интуиционисты» (см. стр. 108) выдвинули свои возражения; некоторые из них даже настаивали, чтобы подобные теоремы были вообще изгнаны из математики. Изучающий математику не должен, впрочем, принимать эти возражения более серьезно, чем это сделало большинство критиков.

***4. Теорема о последовательностях. Компактные множества.**

Пусть x_1, x_2, x_3, ... есть некоторая бесконечная последовательность чисел, различных или нет, содержащихся в *отрезке* I, $a \leqslant x \leqslant b$. Последовательность может стремиться или не стремиться к пределу. Но как бы то ни было, *всегда можно извлечь из такой последовательности, выбрасывая некоторые из ее членов, такую новую бесконечную последовательность* y_1, y_2, y_3, ..., *которая стремилась бы к пределу, заключенному в промежутке* I.

Чтобы доказать эту теорему, разделим интервал I с помощью средней точки $x = \dfrac{a+b}{2}$ на два замкнутых отрезка I' и I'':

$$I'\colon\ a \leqslant x \leqslant \frac{a+b}{2},$$

$$I''\colon\ \frac{a+b}{2} \leqslant x \leqslant b.$$

По крайней мере в одном из них будет находиться бесчисленное количество членов x_n основной последовательности; обозначим его через I_1. Выберем один из этих членов x_{n_1} и обозначим его через y_1. Проделаем то же самое с промежутком I_1. Так как в интервале I_1 имеется бесконечное множество членов x_n, то их должно быть бесконечное множество также и по крайней мере в одной из половин I_1; обозначим эту половину через I_2. На отрезке I_2 возьмем член x_n, для которого $n > n_1$, и обозначим его через y_2. Продолжая таким же образом, мы можем найти последовательность вложенных отрезков I_1, I_2, I_3, ... и подпоследовательность y_1, y_2, y_3, ... членов основной последовательности таким образом, что y_n лежит в интервале I_n, каково бы ни было n. Эта последовательность интервалов стягивается к некоторой точке y промежутка, и ясно, что последовательность y_1, y_2, y_3, ... имеет предел y, что и требовалось доказать.

* Эти рассуждения допускают обобщение того типа, который характерен для современной математики. Рассмотрим переменное X, пробегающее некоторое множество S, в котором каким-то образом определено понятие «расстояния». S может быть множеством точек на плоскости или в пространстве. Но это не является необходимым; например, S может быть также множеством всех треугольников на плоскости. Если X и Y являются двумя треугольниками с вершинами A, B, C и A', B', C' соответственно, то в качестве «расстояния» между треугольниками можно взять, например, число

$$d(X, Y) = AA' + BB' + CC',$$

где AA' обозначает обычное расстояние между точками A и A', и т. д. Как только во множестве введено понятие «расстояния», мы имеем возможность определить понятие последовательности элементов X_1, X_2, X_3, ..., стремящейся к пределу X — также элементу множества S. Мы подразумеваем под

этим, что $d(X, X_n) \to 0$ при $n \to \infty$. Теперь мы скажем, что множество S *компактно, если из каждой последовательности* $X_1, X_2, X_3, \ldots$ *элементов этого множества можно извлечь подпоследовательность, стремящуюся к некоторому пределу* X, *принадлежащему множеству* S. В предыдущем пункте мы показали, что замкнутый промежуток $a \leqslant x \leqslant b$ компактен в указанном смысле. Таким образом, понятие компактного множества можно считать обобщением понятия *замкнутого интервала* на числовой оси. Отметим, что числовая ось *в целом* некомпактна, поскольку последовательность целых чисел 1, 2, 3, 4, 5, ... не стремится ни к какому пределу и не содержит в себе никакой подпоследовательности, которая стремилась бы к пределу. Также и открытый интервал некомпактен, например, $0 < x < 1$, не включающий конечных точек; действительно, последовательность $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ или любая ее подпоследовательность стремится к пределу 0, который не принадлежит, однако, рассматриваемому открытому промежутку. Таким же образом можно показать, что область плоскости, состоящая, скажем, из внутренних точек некоторого квадрата или прямоугольника, некомпактна; но она становится компактной после присоединения точек границы. Нетрудно также убедиться, что множество всех треугольников с вершинами, лежащими внутри или на окружности данного круга, компактно.

Понятие непрерывности допускает обобщение на случай, когда переменное X пробегает любое множество S, лишь бы в этом последнем было предварительно введено понятие стремления к пределу. Говорят, что функция $u = F(X)$ (где u мыслится как действительное число) непрерывна на элементе X, если всякий раз, как последовательность элементов $X_1, X_2, X_3, \ldots$ имеет предел X, соответствующая последовательность чисел $F(X_1), F(X_2), F(X_3), \ldots$ имеет предел $F(X)$. (Можно дать определение и с помощью ε, δ.) Легко также убедиться, что теорема Вейерштрасса остается в силе для случая обобщенной непрерывной функции $F(X)$, заданной на некотором компактном множестве:

Если $u = F(X)$ *есть непрерывная функция, определенная для всех элементов компактного множества* S, *то существует обязательно такой элемент* S, *для которого* $F(X)$ *достигает своего наибольшего значения, и другой элемент, для которого* $F(X)$ *достигает своего наименьшего значения.*

Доказательство не представит никакого труда для того, кто схватил общий характер относящихся сюда идей; мы не пойдем дальше в этом же направлении. Мы увидим в главе VIII, что теорема Вейерштрасса в ее общей формулировке имеет особенно большое значение в теории максимумов и минимумов.

§ 6. Некоторые применения теоремы Больцано

1. Геометрические применения. С помощью простой и общей теоремы Больцано можно доказать некоторые утверждения, на первый взгляд отнюдь не представляющиеся вполне очевидными. Установим, прежде всего, следующее: *если* A *и* B *— две заданные фигуры на плоскости, то существует такая прямая в этой плоскости, которая обе фигуры одновременно делит на равновеликие (в смысле площади) части.* Под «фигурой» здесь понимается всякая часть плоскости, ограниченная

простой замкнутой кривой.

Начнем доказательство с того, что выберем произвольную фиксированную точку P в нашей плоскости и проведем из нее фиксированный луч PR, от которого будем вести отсчет углов. Каков бы ни был луч PS, составляющий угол x с лучом PR, существует направленная прямая, параллельная PS и *делящая фигуру A на равновеликие части.* Действительно, возьмем одну из направленных прямых, параллельных PS, такую что вся фигура A лежит по одну ее сторону; пусть эта прямая будет l_1. Станем подвергать l_1 параллельному переносу таким образом, чтобы при окончательном положении (которое назовем l_2) вся фигура A оказалась уже по другую ее сторону (рис. 173). В таком случае функция, определяемая как разность площади части A, расположенной вправо от направленной прямой, и площади части A, расположенной влево («вправо» — «к востоку», «влево» — «к западу», если прямая направлена, скажем, «на север»), оказывается положительной для положения прямой l_1 и отрицательной для положения l_2. Так как эта функция непрерывна, то, по теореме Больцано, она обращается в нуль при каком-то промежуточном положении прямой, которое мы обозначим теперь через l_x и при котором, очевидно, фигура A разбивается пополам. Итак, каково бы ни было x $(0° \leqslant x < 360°)$, существует прямая l_x, разбивающая A пополам.

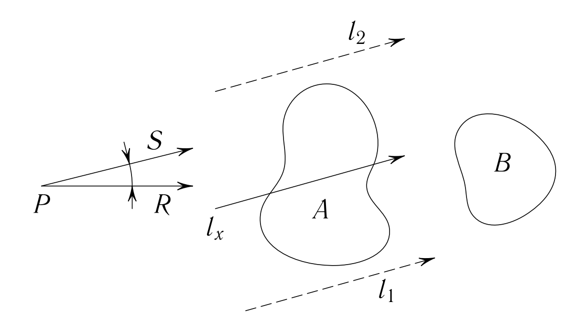


Рис. 173. Одновременное деление пополам двух площадей

Обозначим теперь через $y = f(x)$ разность между площадью части фигуры B справа от l_x и площадью части B слева от l_x. Допустим для определенности, что прямая l_0, параллельная PR и разбивающая A пополам, справа имеет бо́льшую часть площади B, чем слева; тогда y положительно при $x = 0°$. Пусть теперь x возрастает до $180°$; тогда прямая l_{180}, параллельная PR и разбивающая A пополам, совпадает с l_0 (но направлена в противоположную сторону, а «правая» и «левая» стороны переместились); отсюда ясно, что значение y при $x = 180°$ численно то же, что и при $x = 0°$, но с обратным знаком, т. е. отрицательно. Так как y есть функция x, непрерывная при $0° \leqslant x \leqslant 180°$ (упомянутая

разность площадей, очевидно, изменяется непрерывно при вращении секущей прямой), то существует такое значение $x = \alpha$, при котором y обращается в нуль. Но тогда прямая l_α разбивает пополам обе фигуры A и B одновременно. Наша теорема доказана.

Следует заметить, что мы установили всего-навсего *существование* прямой, обладающей заданным свойством, но не указали определенной процедуры для ее *построения*: в этом — характерная черта «чистых» математических доказательств существования.

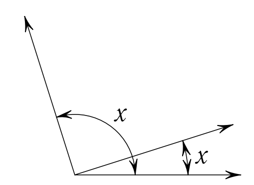


Рис. 174. Деление площади на четыре равные части

Вот другая, аналогичная проблема: дана одна фигура A на плоскости; требуется разбить ее на *четыре* равновеликие части *двумя взаимно перпендикулярными* прямыми. Чтобы доказать существование решения, вернемся к тому этапу решения предыдущей проблемы, когда была введена прямая l_x, но фигура B еще не была введена в рассуждение. Рассмотрим прямую l_{x+90}, перпендикулярную к l_x и также разбивающую A пополам. Если занумеруем четыре части A так, как показано на рис. 174, то получим, очевидно,

$$A_1 + A_2 = A_3 + A_4$$

и

$$A_2 + A_3 = A_1 + A_4.$$

Отсюда вычитанием получим

$$A_1 - A_3 = A_3 - A_1,$$

т. е.

$$A_1 = A_3,$$

а значит,

$$A_2 = A_4.$$

Итак, существование решения нашей проблемы будет доказано, если установим существование такого угла α, что для прямой l_α будет удо-

влетворено равенство двух частей нашей фигуры

$$A_1(\alpha) = A_2(\alpha),$$

так как отсюда будет вытекать равенство всех четырех частей. Рассмотрим теперь функцию $y = f(x)$,

$$f(x) = A_1(x) - A_2(x),$$

где $A_1(x)$ и $A_2(x)$ — части фигуры, соответствующие прямой l_x. При $x = 0^\circ$ пусть будет, например, $f(0) = A_1(0) - A_2(0) > 0$. Тогда при $x = 90^\circ$ получится: $f(90) = A_1(90) - A_2(90) = A_2(0) - A_3(0) = A_2(0) - A_1(0) < 0$. Но $f(x)$ — непрерывная функция; значит, при каком-то значении α между 0° и 90° получится $f(\alpha) = A_1(\alpha) - A_2(\alpha) = 0$. Тогда прямые l_α и $l_{\alpha+90}$ разбивают фигуру на четыре равновеликие части.

Эти проблемы обобщаются на случай трех и большего числа измерений. В случае трех измерений первая проблема формулируется следующим образом: даны три пространственных тела; требуется найти плоскость, разбивающую каждое из них пополам одновременно. Доказательство возможности решения также основывается на теореме Больцано. В случае большего числа измерений аналогичное утверждение также справедливо, но доказательство требует применения более тонких методов.

***2. Применение к одной механической проблеме.** Мы закончим эту главу рассмотрением одной на первый взгляд трудной механической проблемы, которая, однако, решается очень просто посредством соображений, связанных с непрерывностью. (Проблема была предложена Г. Уитни.)

Предположим, что поезд на протяжении некоторого конечного промежутка времени проходит прямолинейный отрезок пути от станции A до станции B. Вовсе не предполагается, что движение происходит с постоянной скоростью или с постоянным ускорением. Напротив, поезд может двигаться как угодно: с ускорениями, с замедлениями; не исключены даже мгновенные остановки или частично даже движение в обратном направлении, прежде чем в итоге поезд придет на станцию B. Но так или иначе движение поезда на протяжении всего временно́го промежутка считается известным заранее; другими словами, считается заданной функция $s = f(t)$, где s — расстояние поезда от станции A, а t — время, отсчитываемое от момента отправления поезда. К полу одного из вагонов прикреплен на шарнире твердый тяжелый стержень, который без трения может двигаться вокруг оси, параллельной осям вагонов, вперед и назад — от пола до пола. (Мы допускаем, что, прикоснувшись к полу, он в дальнейшем останется на нем лежать, если ему не случится «подпрыгнуть» снова.) Вопрос заключается в следующем: *возможно ли в момент отхода поезда поместить стержень в такое начальное*

положение, т. е. дать ему такой угол наклона, чтобы на протяжении всего пути он не прикоснулся к полу, будучи предоставлен воздействию движения поезда и силе собственной тяжести?

На первый взгляд может показаться совершенно невероятным, чтобы при наперед определенной схеме движения поезда взаимодействие силы тяжести и сил реакции было способно обеспечить требуемое равновесие стержня при единственном условии — надлежащем выборе начального положения. Но мы сейчас установим, что такое начальное положение всегда существует.

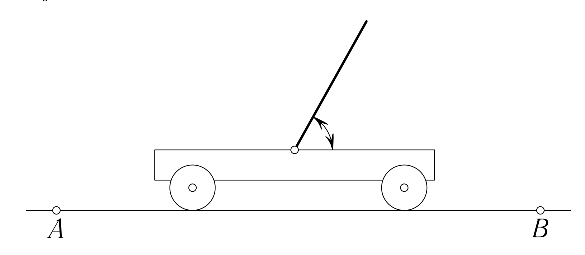


Рис. 175. Проблема Уитни

К счастью, доказательство не подразумевает точного знания законов динамики. (Исходя из этих законов, решить задачу было бы чрезвычайно трудно.) Достаточно принять только одно допущение физического содержания: *последующее движение стержня зависит непрерывно от его начального положения*; в частности, если при данном начальном положении стержень во время пути упадет на пол в одну из сторон, то при всяком начальном положении, достаточно мало отличающемся от данного, он не упадет на пол в противоположную сторону.

Обратим теперь внимание на то, что во всякий момент времени положение стержня характеризуется углом α, который он составляет с полом. Углам $\alpha = 0^\circ$ и $\alpha = 180^\circ$ соответствуют два взаимно противоположных горизонтальных (лежачих) положения. Обозначим через x значение угла α в начальном положении стержня. Доказательство нашего утверждения будет косвенное, в соответствии с чисто экзистенциальным характером самой проблемы. Допустим, что *всегда*, т. е. при любом начальном положении стержня, стержень непременно упадет или в одну, или в другую сторону, так что α примет значение или 0°, или 180°. Определим тогда функцию $f(x)$ так: $f(x) = +1$ или -1 смотря по тому, упадет ли стержень в сторону, соответствующую углу $\alpha = 0^\circ$ или углу $\alpha = 180^\circ$. Свойства функции $f(x)$ таковы: она задана в интервале $0 \leqslant x \leqslant 180^\circ$, непрерывна в нем, и притом $f(0) = +1$, $f(180) = -1$. Отсюда, по теореме Больцано, следует, что при каком-то промежуточном значении x ($0^\circ < x < 180^\circ$) должно выполняться равенство $f(x) = 0$. А это противоречит тому, что функция $f(x)$ может принимать только значения $+1$ и -1.

Значит, приходится отвергнуть сделанное допущение, согласно которому стержень упадет на пол во время движения поезда при каком угодно начальном его положении.

Совершенно ясно, что приведенное доказательство носит чисто теоретический характер, потому что не дает решительно никаких указаний на то, как определить искомое начальное положение стержня. Вместе с тем, даже если бы такое положение и могло быть вычислено теоретически с абсолютной точностью, практически оно было бы бесполезно вследствие своей неустойчивости. Так, например, в предельном случае, если поезд неподвижен в течение всего «путешествия», решение совершенно очевидно: $x = 90°$; но всякий, кто пытался уравновесить иголку в стоячем положении на гладкой горизонтальной поверхности, понимает, насколько это решение практически нереально. Тем не менее с математической точки зрения приведенное доказательство имеет неоспоримый интерес.

Упражнения. 1) Обобщите это рассуждение на случай, когда «путешествие» продолжается бесконечно долго.[1]

2) Обобщите также на случай, когда поезд движется по произвольной кривой на плоскости, а стержень может падать в любом направлении, т. е. стержень обладает двумя степенями свободы. (*Указание*: невозможно непрерывно отобразить круговой диск на одну его окружность, оставляя все точки окружности неподвижными — см. стр. 272.)

ДОПОЛНЕНИЕ К ГЛАВЕ VI

Дальнейшие примеры на пределы и непрерывность

§ 1. Примеры пределов

1. Общие замечания. Во многих случаях сходимость последовательности a_n может быть доказана по следующей схеме. Мы рассматриваем две другие последовательности b_n и c_n, члены которых, вообще говоря, имеют более простую структуру и обладают тем свойством, что

$$b_n \leqslant a_n \leqslant c_n \tag{1}$$

[1] Например, воспользуйтесь принципом стягивающихся отрезков (см. стр. 88); если из какого-то положения стержень не упадет за фиксированное конечное время, то за это же время он не успеет упасть также из всех достаточно близких положений. — *Прим. ред. наст. изд.*

при всех значениях n. Тогда, *если будет установлено, что последовательности* b_n *и* c_n *имеют один и тот же предел* α, *можно будет утверждать, что последовательность* a_n *также имеет предел* α. Формальное доказательство этой теоремы мы можем предоставить читателю.

Ясно, что применение указанной схемы потребует оперирования неравенствами. В связи с этим своевременно напомнить небольшое число элементарных правил, которым подчинены арифметические операции с неравенствами.

1. Если $a > b$, то $a + c > b + c$ (к обеим частям неравенства можно прибавить одно и то же число).

2. Если $a > b$ и c — *положительно*, то $ac > bc$ (можно умножить неравенство на положительное число).

3. Если $a > b$, то $-b > -a$ (направление неравенства меняется при умножении на -1). Так, $2 < 3$, но $-3 < -2$.

4. Если a и b одного и того же знака, то из неравенства $a < b$ следует $\frac{1}{a} > \frac{1}{b}$.

5. $|a + b| \leqslant |a| + |b|$.

2. Предел q^n. Если q — число большее чем 1, то члены последовательности q^n неограниченно возрастают; например, так будет при $q = 2$:

$$q,\ q^2,\ q^3,\ \ldots$$

Такие последовательности «стремятся к бесконечности» (см. стр. 316). В самом общем случае доказательство этого основывается на важном неравенстве (см. стр. 34)

$$(1 + h)^n \geqslant 1 + nh \geqslant nh, \tag{2}$$

где h — какое угодно положительное число. Мы положим $q = 1 + h$, где $h > 0$; тогда

$$q^n = (1 + h)^n > nh.$$

Пусть k — сколь угодно большое положительное число; в таком случае достаточно взять $n > \frac{k}{h}$, чтобы получить неравенство

$$q^n > nh > k;$$

значит, $q^n \to \infty$. Если $q = 1$, то все члены последовательности равны 1, и значит, предел последовательности есть 1. Если q отрицательно, то знаки q^n чередуются, и в случае $|q| \geqslant 1$ предела нет.

Упражнение. Дайте строгое доказательство последнему утверждению.

На стр. 84 мы установили, что если $-1 < q < 1$, то $q^n \to 0$. Дадим здесь еще другое, очень простое доказательство. Рассмотрим случай $0 < q < 1$. Тогда члены последовательности q, q^2, q^3, … монотонно убывают,

оставаясь положительными. Отсюда следует (см. стр. 317), что последовательность имеет предел: $q^n \to a$. Умножая обе части последнего соотношения на q, получим: $q^{n+1} \to aq$.

Но q^{n+1} должно иметь тот же предел, что и q^n, так как дело не меняется от того, как обозначен возрастающий показатель — через n или через $n+1$. Значит, $aq = a$, или $a(q-1) = 0$. Так как по условию $(1-q) \neq 0$, то отсюда следует $a = 0$.

Если $q = 0$, предыдущее утверждение тривиально. Если, наконец, $-1 < q < 0$, то $0 < |q| < 1$; поэтому, как только что доказано, $|q^n| = |q|^n \to 0$. Но в таком случае $q^n \to 0$ при условии $|q| < 1$. Доказательство закончено.

Упражнения. Докажите, что при $n \to \infty$

1) $\left(\dfrac{x^2}{1+x^2}\right)^n \to 0,$

2) $\left(\dfrac{x}{1+x^2}\right)^n \to 0,$

3) $\left(\dfrac{x^3}{4+x^2}\right)^n$ стремится к бесконечности при $x > 2$, к нулю при $|x| < 2$.

3. Предел $\sqrt[n]{p}$. Последовательность чисел

$$a_n = \sqrt[n]{p},$$

т. е.

$$p, \quad \sqrt{p}, \quad \sqrt[2]{p}, \quad \sqrt[4]{p}, \quad \ldots$$

имеет предел 1, каково бы ни было положительное число p:

$$\sqrt[n]{p} \to 1 \quad \text{при} \quad n \to \infty. \tag{3}$$

(Символ $\sqrt[n]{p}$ обозначает, как всегда, положительный корень степени n. В случае, если p отрицательно, при n четном не существует действительных корней степени n.)

Докажем соотношение (3). Предположим прежде всего, что $p > 1$; тогда также $\sqrt{p} > 1$. Мы можем положить

$$\sqrt[n]{p} = 1 + h_n,$$

причем h_n — положительная величина, зависящая от n. Из неравенства (2) следует

$$p = (1 + h_n)^n > nh_n.$$

Деление на n дает

$$0 < h_n < \frac{p}{n}.$$

Так как последовательности $b_n = 0$ и $c_n = \dfrac{p}{n}$ обе имеют предел 0, то на основании рассуждения, приведенного в пункте 1, h_n также при возрастании n имеет предел 0, и наше утверждение, таким образом, доказано в случае $p > 1$. Мы встретились здесь с очень типическим примером, когда предельное соотношение, в данном случае $h_n \to 0$, устанавливается

посредством заключения h_n между двумя границами, пределы которых определяются более просто.

Кстати, мы получили оценку для разности h_n между $\sqrt[n]{p}$ и 1: эта разность непременно меньше, чем $\frac{p}{n}$.

Если $0 < p < 1$, то $\sqrt[n]{p} < 1$, и можно положить

$$\sqrt[n]{p} = \frac{1}{1+h_n},$$

где h_n — снова положительное число, зависящее от n. Отсюда следует

$$p = \frac{1}{(1+h_n)^n} < \frac{1}{nh_n},$$

так что

$$0 < h_n < \frac{1}{np},$$

и, значит, h_n стремится к нулю при $n \to \infty$. И тогда, очевидно, $\sqrt[n]{p} \to 1$.

«Уравнивающее» воздействие извлечения корня степени n, выражающееся в том, что результаты извлечения корней последовательно возрастающих степеней из данного положительного числа приближаются к единице, остается в силе иногда и в том случае, если само подрадикальное выражение не остается постоянным. Мы проверим сейчас, что

$$\sqrt[n]{n} \to 1 \quad \text{при} \quad n \to \infty.$$

Небольшое ухищрение позволит нам сослаться опять на неравенство (2). Вместо корня степени n из n возьмем корень степени n из $\sqrt{n}$. Полагая $\sqrt[n]{\sqrt{n}} = 1 + k_n$, где k_n — положительная величина, зависящая от n, получим с помощью упомянутого неравенства $\sqrt{n} = (1+k_n)^n > nk_n$, так что

$$k_n < \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}.$$

Значит,

$$1 < \sqrt[n]{n} = (1+k_n)^2 = 1 + 2k_n + k_n^2 < 1 + \frac{2}{\sqrt{n}} + \frac{1}{n}.$$

Правая часть этого неравенства стремится к 1 при $n \to \infty$, и потому то же самое можно сказать относительно $\sqrt[n]{n}$.

4. Разрывные функции как предел непрерывных. Будем рассматривать такие последовательности a_n, в которых члены a_n — не постоянные числа, а функции некоторой переменной x, именно $a_n = f_n(x)$. Если только такая последовательность сходящаяся, то ее предел также есть функция x:

$$f(x) = \lim f_n(x).$$

Такого рода представления функции $f(x)$ в виде предела других функций часто бывают полезны, так как «более сложные» функции таким образом приводятся к «более простым».

В частности, это обнаруживается при рассмотрении некоторых явных формул, определяющих функции с разрывами. Рассмотрим, например, последовательность $f_n(x) = \frac{1}{1+x^{2n}}$. При $|x| = 1$ мы получаем $x^{2n} = 1$, $f_n(x) = \frac{1}{2}$, так что $f_n(x) \to \frac{1}{2}$. С другой стороны, при $|x| < 1$ мы имеем $x^{2n} \to 0$ и $f_n(x) \to 1$; наконец, при $|x| > 1$ получается $x^{2n} \to \infty$ и, следовательно, $f_n(x) \to 0$. В итоге

$$f(x) = \lim \frac{1}{1+x^{2n}} = \begin{cases} 1 & \text{при } |x| < 1, \\ \frac{1}{2} & \text{при } |x| = 1, \\ 0 & \text{при } |x| > 1. \end{cases}$$

Мы видим, что разрывная функция $f(x)$ представлена как предел последовательности непрерывных рациональных функций.

Другой интересный пример в таком же роде дается последовательностью

$$f_n(x) = x^2 + \frac{x^2}{1+x^2} + \frac{x^2}{(1+x^2)^2} + \ldots + \frac{x^2}{(1+x^2)^n}.$$

При $x = 0$ все функции $f_n(x)$ обращаются в нуль, и потому $f(0) = \lim f_n(0) = 0$. При $x \neq 0$ выражение $\frac{1}{1+x^2}$ положительно и меньше чем 1, и потому теория геометрической прогрессии позволяет утверждать, что $f_n(x)$ сходится при $n \to \infty$. Предел, т. е. сумма бесконечной прогрессии, равен $\frac{x^2}{1-q} = \frac{x^2}{1 - \frac{1}{1+x^2}}$, т. е. $1 + x^2$. Итак, $f_n(x)$ стремится к функции $f(x) = 1 + x^2$ при $x \neq 0$ и к $f(x) = 0$ при $x = 0$. Получается функция $f(x)$ с устранимым разрывом в точке $x = 0$.

***5. Пределы при итерации.** Нередко приходится рассматривать последовательности, сконструированные таким образом, что a_{n+1} получается из a_n посредством тех же операций, посредством каких a_n получается из a_{n-1}: эта процедура позволяет вычислить любой член последовательности, если известен первый. В подобных случаях говорят о процедуре «итерации».

Примером может служить последовательность

$$1,\ \sqrt{1+1},\ \sqrt{1+\sqrt{2}},\ \sqrt{1+\sqrt{1+\sqrt{2}}},\ \ldots;$$

каждый член ее получается из предыдущего посредством прибавления единицы и затем извлечения квадратного корня. Таким образом, последовательность определяется соотношениями

$$a_1 = 1, \qquad a_{n+1} = \sqrt{1 + a_n}.$$

Найдем ее предел. Очевидно, что при $n > 1$ имеем $a_n > 1$. Далее, последовательность наша монотонно возрастает, так как

$$a_{n+1}^2 - a_n^2 = (1 + a_n) - (1 + a_{n-1}) = a_n - a_{n-1}.$$

Раз только $a_n > a_{n-1}$, то значит, $a_{n+1} > a_n$; но $a_2 - a_1 = \sqrt{2} - 1 > 0$, и потому (с помощью индукции) отсюда следует, что $a_{n+1} > a_n$ при всех значениях n. Мы замечаем дальше, что рассматриваемая последовательность ограниченная; в самом деле,

$$a_{n+1} = \frac{1 + a_n}{a_{n+1}} < \frac{1 + a_{n+1}}{a_{n+1}} = 1 + \frac{1}{a_{n+1}} < 2.$$

В силу принципа монотонных последовательностей отсюда вытекает существование предела: $a_n \to a$, причем $1 < a \leqslant 2$. Но ясно видно, что a есть положительный корень уравнения $x^2 = 1 + x$: действительно, соотношение $a_{n+1}^2 = 1 + a_n$ в пределе при $n \to \infty$ дает нам $a^2 = 1 + a$. Решая уравнение, мы убеждаемся, что $a = \dfrac{1 + \sqrt{5}}{2}$.

Отсюда, наоборот, легко сделать тот вывод, что это квадратное уравнение можно решать приближенно, с какой угодно степенью точности, посредством итерационной процедуры.

Таким же образом, с помощью итераций, можно решать и другие алгебраические уравнения. Например, кубическое уравнение $x^3 - 3x + 1 = 0$ напишем в форме

$$x = \frac{1}{3 - x^2}$$

и затем, взяв в качестве a_1 произвольное число, скажем $a_1 = 0$, определим дальше последовательность a_n по формуле

$$a_{n+1} = \frac{1}{3 - a_n^2};$$

при этом получим

$$a_2 = \frac{1}{3} = 0{,}3333\ldots, \qquad a_3 = \frac{9}{26} = 0{,}3461\ldots, \qquad a_4 = \frac{676}{1947} = 0{,}3472\ldots$$

Можно показать, что последовательность имеет предел, равный корню данного кубического уравнения, а именно $a = 0{,}3473\ldots$

Итерационные процессы в этом роде имеют громадное значение в математике, так как с их помощью большей частью даются «доказательства существования»; они полезны и в приложениях, где доставляют методы приближенного решения разнообразных проблем.

Упражнения на пределы. При $n \to \infty$:

1) Докажите, что $\sqrt{n+1} - \sqrt{n} \to 0$. $\Big($*Указание*: напишите разность в виде $\dfrac{\sqrt{n+1} - \sqrt{n}}{\sqrt{n+1} + \sqrt{n}} \cdot (\sqrt{n+1} + \sqrt{n})$.$\Big)$

2) Найдите предел $\sqrt{n^2 + a} - \sqrt{n^2 + b}$.
3) Найдите предел $\sqrt{n^2 + an + b} - n$.
4) Найдите предел $\dfrac{1}{\sqrt{n+1} + \sqrt{n}}$.
5) Докажите, что предел $\sqrt[n]{n+1}$ равен 1.
6) Каков предел $\sqrt[n]{a^n + b^n}$, если $a > b > 0$?

7) Каков предел $\sqrt[n]{a^n + b^n + c^n}$, если $a > b > c > 0$?

8) Каков предел $\sqrt[n]{a^n b^n + a^n c^n + b^n c^n}$, если $a > b > c > 0$?

9) Мы увидим позднее (стр. 473), что $e = \lim\left(1 + \dfrac{1}{n}\right)^n$. Используя это, найдите предел $\lim\left(1 + \dfrac{1}{n^2}\right)^n$.

§ 2. Пример, относящийся к непрерывности

Чтобы дать формальное доказательство непрерывности данной функции, требуется проверка согласно определению, приведенному на стр. 332. Иногда соответствующая процедура оказывается очень громоздкой, но, к счастью, мы имеем право сослаться на обстоятельство, которое будет установлено в главе VIII, а именно: непрерывность следует из дифференцируемости. Так как дифференцируемость там же будет установлена систематически для всех элементарных функций, то мы (как это обычно делается) воздержимся от того, чтобы приводить скучные доказательства непрерывности функций различных типов.

Но в качестве дальнейшей иллюстрации общего определения мы все же рассмотрим здесь еще один пример, именно функцию

$$f(x) = \frac{1}{1 + x^2}.$$

Мы имеем право ограничить возможные изменения x конечным интервалом $|x| \leqslant M$, где, впрочем, M как угодно велико. Написав

$$f(x_1) - f(x) = \frac{1}{1 + x_1^2} - \frac{1}{1 + x^2} = \frac{x^2 - x_1^2}{(1 + x^2)(1 + x_1^2)} = (x - x_1)\frac{(x + x_1)}{(1 + x^2)(1 + x_1^2)},$$

мы видим, что при $|x| \leqslant M$ будет и $|x_1| \leqslant M$. Отсюда следует неравенство

$$|f(x_1) - f(x)| \leqslant |x - x_1| \cdot |x + x_1| \leqslant |x - x_1| \cdot 2M.$$

Значит, разность в левой части станет меньше, чем наперед заданное положительное число ε, при условии, что будет

$$|x_1 - x| < \delta = \frac{\varepsilon}{2M}.$$

Следует отметить, что мы были слишком великодушны в этих оценках. Читатель без труда убедится, что при больших значениях x и x_1 можно было бы удовлетвориться гораздо большими значениями δ.

ГЛАВА VII

Максимумы и минимумы

Введение

Отрезок прямой линии определяет кратчайший путь между двумя его конечными точками. Дуга большого круга представляет собой кратчайшую кривую, которой можно соединить две точки на сфере. Среди всех замкнутых плоских кривых одной и той же длины наибольшая площадь охватывается окружностью, а среди всех замкнутых поверхностей одной и той же площади наибольший объем охватывается сферой.

Максимальные и минимальные свойства подобного рода были известны еще древнегреческим математикам, хотя и не всегда со строгими их доказательствами. Одно из самых замечательных относящихся сюда открытий приписывается Герону, александрийскому ученому I столетия нашей эры. Издавна было известно, что световой луч, выходящий из точки P и встречающийся с плоским зеркалом L, отражается в направлении некоторой точки Q таким образом, что PR и QR образуют одинаковые углы с зеркалом. По преданию, Герон установил, что если R' — любая точка зеркала, отличная от R, то сумма отрезков $PR' + R'Q$ больше, чем $PR + RQ$. Эта теорема (которую мы скоро докажем) характеризует истинный путь светового луча PRQ между P и Q как кратчайший путь от P к Q с заходом на зеркало L — открытие, которое можно рассматривать как зародыш теории геометрической оптики.

Нет ничего удивительного в том, что математики живейшим образом интересуются подобного рода вопросами. В повседневной жизни постоянно возникают проблемы наибольшего и наименьшего, наилучшего и наихудшего. Именно в такой форме могут быть поставлены многие задачи, имеющие практическое значение. Например, каковы должны быть очертания судна, для того чтобы оно испытывало при движении в воде наименьшее сопротивление? Каково должно быть соотношение размеров цилиндрического резервуара, чтобы при заданном расходе материала объем был наибольшим?

Возникнув в XVII столетии, общая теория экстремальных, т. е. максимальных и минимальных, значений величин выдвинула обширный ряд принципов науки, служащих целям обобщения и систематизации.

Первые шаги, сделанные Ферма в области дифференциального исчисления, были ускорены стремлением найти общие методы для изучения вопросов о максимумах и минимумах. В последующем столетии эти методы были значительно обогащены с изобретением вариационного исчисления. Становилось все яснее и яснее, что физические законы природы в высшей степени удачно формулируются в терминах принципа минимальности, обеспечивающего естественный подход к более или менее полному решению частных проблем. Одним из самых замечательных достижений современной математики является теория стационарных значений, дающая такого рода расширение понятия максимума и минимума, которое базируется одновременно на анализе и на топологии.

Мы будем здесь рассматривать весь вопрос в целом с совершенно элементарной точки зрения.

§ 1. Задачи из области элементарной геометрии

1. Треугольник наибольшей площади при двух заданных сторонах. Даны два отрезка a и b; требуется найти треугольник возможно большей площади, у которого две стороны были бы a и b. Решением является *прямоугольный* треугольник с катетами a и b. Рассмотрим в самом деле какой-нибудь треугольник с двумя сторонами a и b (рис. 176). Если h есть высота, соответствующая основанию a, то площадь треугольника A равна $\frac{1}{2}ah$. Это последнее выражение, очевидно, принимает наибольшее значение при наибольшем возможном значении h, что случится именно при h, равном b, т. е. тогда, когда треугольник прямоугольный. Итак, максимальная площадь равна $\frac{1}{2}ab$.

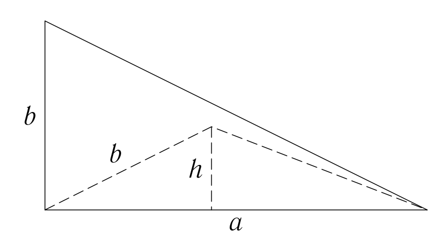


Рис. 176. Максимум площади треугольника при двух данных сторонах

2. Теорема Герона. Экстремальное свойство световых лучей. Дана прямая L и две точки P и Q по одну и ту же ее сторону. Как выбрать точку R на прямой L с таким расчетом, чтобы сумма отрезков $PR + RQ$ давала кратчайший путь от P к Q с заходом на L? В этом заключается проблема Герона о световом луче (точно такую же проблему приходится решать тому, кто, желая из точки P как можно скорее пройти в точку Q, должен был бы по дороге подойти к L: представьте себе, что L — берег реки, и там нужно зачерпнуть ведро воды). Чтобы получить решение, построим зеркальное отраже-

ние P' точки P относительно прямой L, и тогда прямая $P'Q$ пересекает L как раз в искомой точке R. Легко доказать, что $PR + RQ$ меньше, чем $PR' + R'Q$, где R' — любая точка на L, отличная от R. Действительно, $PR = P'R$ и $PR' = P'R'$, значит, $PR + RQ = P'R + RQ = P'Q$ и $PR' + R'Q = P'R' + R'Q$. Но $P'R' + R'Q$ больше, чем $P'Q$ (так как сумма двух сторон треугольника больше третьей стороны), т. е. $PR' + R'Q$ больше, чем $PR + RQ$, что и требовалось доказать. В дальнейшем существенно предполагать, что P и Q не лежат на самой прямой L.

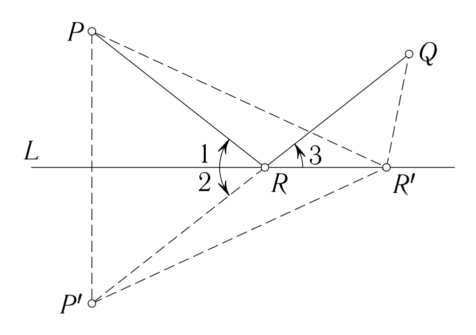


Рис. 177. Теорема Герона

Из рис. 177 видно, что $\angle 3 = \angle 2$ и $\angle 2 = \angle 1$, так что $\angle 1 = \angle 3$. Другими словами, точка R такова, что PR и QR образуют одинаковые углы с L. Отсюда следует, что световой луч, отражающийся от L (а при отражении, как показывает эксперимент, угол падения равен углу отражения), действительно обращает в минимум путь из P в Q с заходом на L — в согласии с высказанным утверждением.

Задачу можно обобщить, вводя несколько прямых L, M, ... Рассмотрим, например, случай, когда имеются две прямые L, M и две точки P, Q, расположенные, как на рис. 178, и поставим целью найти кратчайший путь из P в Q с заходом сначала на L, потом на M. Пусть Q' — отражение Q относительно M и Q'' — отражение Q' относительно L. Проведем прямую PQ'', пересекающую L в точке R, и прямую RQ', пересекающую M в точке S; тогда $PR + RS + SQ$ и есть искомый кратчайший путь. Доказательство подобно приведенному выше

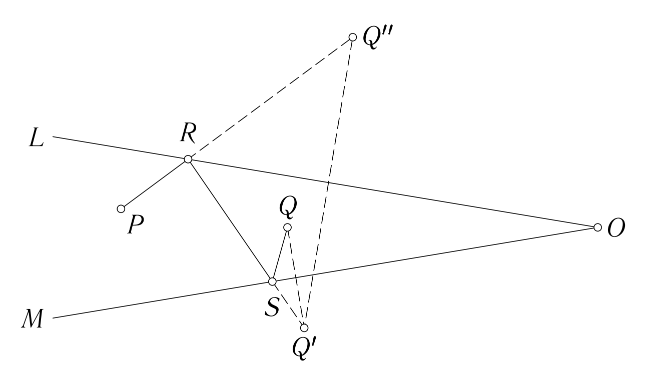


Рис. 178. Отражение в двух зеркалах

и предоставляется читателю в качестве упражнения. Если бы L и M были зеркалами, то световой луч из P, приходящий после отражения в L, потом в M в точку Q, попадал бы на L в точке R, а на M — в точке S; итак, световой луч опять-таки избрал бы для себя путь наименьшей длины.

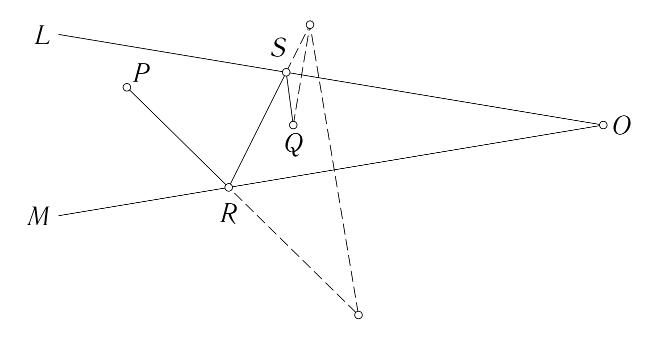


Рис. 179. Вариант предыдущей задачи

Можно было бы также поставить задачу нахождения кратчайшего пути из P в Q с заходом сначала на M, потом на L. Таким должен быть путь $PRSQ$ (рис. 179), определяемый аналогично пути $PRSQ$, рассмотренному раньше. Длина этого нового пути может оказаться или большей, или меньшей, или равной длине прежнего пути.

* **Упражнение.** Покажите, что новый путь больше прежнего в том случае, если точка P и прямая M лежат по одну сторону прямой OQ. В каком случае новый и прежний пути окажутся равными?

3. Применения к задачам о треугольниках. С помощью теоремы Герона можно легко решить следующие две задачи.

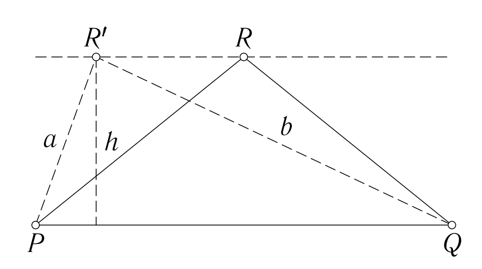


Рис. 180. Треугольник наименьшего периметра при данных основании и площади

а) Задана заранее площадь A и одна сторона $c = PQ$ треугольника; среди всех такого рода треугольников требуется найти тот, для которого сумма двух других сторон a и b наименьшая. Вместо того чтобы задавать сторону c и площадь A треугольника, можно задать сторону c и высоту h, опущенную на c, так как $A = \frac{1}{2}hc$. Таким образом, задача сводится к тому, чтобы найти точку R (рис. 180), находящуюся на расстоянии h от прямой PQ, и притом такую, что сумма сторон $a + b$ обращается в минимум. Из первого условия следует, что точка R долж-

на быть расположена на прямой, параллельной прямой PQ и отстоящей от нее на расстоянии h. Раз это установлено, становится ясно, что задача решается с помощью теоремы Герона в применении к тому случаю, когда P и Q находятся на одном и том же расстоянии от прямой L: искомый треугольник PRQ равнобедренный.

б) Пусть в треугольнике даны одна сторона c и сумма $a + b$ двух других сторон; требуется из всех таких треугольников выбрать тот, у которого площадь наибольшая. Эта задача — обратная по отношению к задаче а). Решением является опять-таки равнобедренный треугольник, для которого $a = b$. Как мы уже видели, для такого треугольника при заданной площади сумма $a + b$ принимает наименьшее значение; это значит, что во всяком другом треугольнике с основанием c и той же площадью сумма $a + b$ имеет большее значение. С другой стороны, из а) ясно, что во всяком треугольнике с основанием c и площадью большей, чем площадь рассматриваемого равнобедренного треугольника, значение $a + b$ также будет больше. Отсюда следует, что всякий другой треугольник, имеющий заданные значения для $a + b$ и для c, должен иметь меньшую площадь, так что наибольшую площадь при заданных c и $a + b$ имеет именно равнобедренный треугольник.

4. Свойства касательных к эллипсу и гиперболе. Соответствующие экстремальные свойства. С теоремой Герона связаны некоторые важные геометрические задачи. Мы установили, что если R — такая точка на прямой L, что $PR + RQ$ обращается в минимум, то прямые PR и RQ образуют одинаковые углы с L. Обозначим минимальное значение $PR + RQ$ через $2a$. Пусть, с другой стороны, p и q обозначают расстояния произвольной точки плоскости соответственно от точек P и Q; рассмотрим геометрическое место всех точек плоскости, для которых $p + q = 2a$. Это геометрическое место — эллипс с фокусами P и Q, проходящий через точку R на прямой L, причем *прямая L касается этого эллипса в точке R*. Действительно, если бы прямая L пересекала эллипс еще в какой-то точке, кроме R, то существовал бы отрезок прямой L, лежащий внутри эллипса; для каждой точки этого отрезка $p + q$ было бы меньше, чем $2a$: в самом деле, легко убедиться, что $p + q$ меньше или больше,

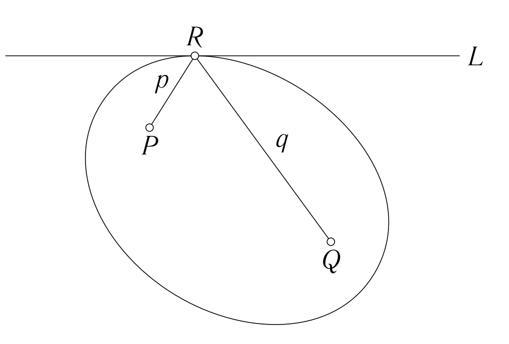


Рис. 181. Свойство касательной к эллипсу

чем $2a$, смотря по тому, находится ли рассматриваемая точка внутри или вне эллипса. Но так как мы знаем, что для точек на прямой L непременно $p+q \geqslant 2a$, то сделанное предположение приходится отбросить. Итак, прямая L — касательная к эллипсу в точке R. Кроме того, мы знаем, что PR и RQ образуют одинаковые углы с L; отсюда в качестве побочного результата наших рассуждений вытекает важная теорема: касательная к эллипсу образует равные углы с прямыми, проведенными из фокусов в точку касания.

Следующая задача родственна предыдущей. Дана прямая линия L и две точки P и Q по *разные* стороны L (рис. 182); требуется найти такую точку R на L, чтобы величина $|p-q|$, т. е. абсолютная величина *разности* расстояний точки R от P и Q, была как можно *больше*. (Мы допускаем, что L не является перпендикуляром, восставленным из середины отрезка PQ: иначе $p-q$ равнялось бы нулю для всякой точки L, и задача потеряла бы смысл.) Приступая к решению задачи, построим зеркальное отражение точки P относительно L: полученная точка P' расположена по ту же сторону L, что и Q. Какова бы ни была точка R' на L, мы имеем: $p = R'P = R'P'$, $q = R'Q$. Так как разность двух сторон треугольника никогда не превышает третьей стороны, то, рассматривая треугольник $R'QP'$, можно заключить, что величина $|p-q| = |R'P' - R'Q|$ меньше или равна $P'Q$; и, как видно из чертежа, только при условии, что R', P' и Q расположены на одной прямой, $|p-q|$ может оказаться *равным* $P'Q$. Поэтому искомая точка R есть точка пересечения прямой L с прямой, проведенной через P' и Q. Как и в предыдущей задаче, не представляет труда установить, ссылаясь на конгруэнтность треугольников RPR' и $RP'R'$, что углы, которые отрезки RP и RQ составляют с прямой L, одинаковы.

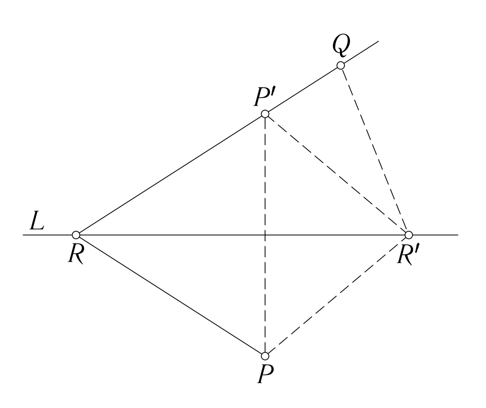


Рис. 182. $|PR - QR| = \text{maximum}$

Отсюда, как и в прежней задаче, уже ничего не стоит получить свойство касательной к гиперболе. Принимая наибольшее значение разности $|PR - RQ|$ равным $2a$, рассмотрим геометрическое место всех точек в плоскости, для которых абсолютная величина $p-q$ равна $2a$. Это — гипербола с фокусами P и Q, проходящая через точку R. Легко убедиться, что абсолютная величина $p-q$ меньше чем $2a$ в области, заключенной между двумя ветвями гиперболы, и больше чем $2a$ по ту сторону каждой из ветвей, по которую лежит соответствую-

щий фокус. Отсюда — в основном с помощью такой же аргументации, как и в случае эллипса, — вытекает, что прямая L касается гиперболы в точке R. К которой именно из ветвей прямая L является касательной, — это зависит от того, которая из точек P и Q ближе к L: если ближе точка P, то касается прямой L та ветвь, которая окружает P; и аналогично для Q (рис. 183). Если P и Q находятся на равных расстояниях от прямой L, то L не касается ни той, ни другой ветви гиперболы, а является одной из ее асимптот. Об этом результате позволительно догадываться исходя из того соображения, что описанное выше построение в рассматриваемом случае не дает никакой (конечной) точки R, так как прямая $P'Q$ оказывается параллельной прямой L.

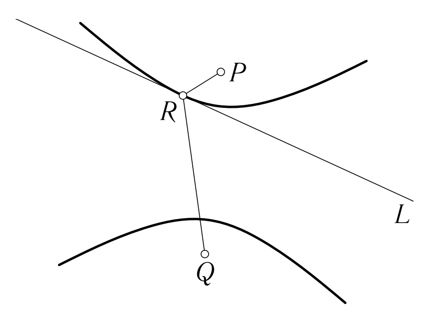


Рис. 183. Свойство касательной к гиперболе

Так же как и в случае эллипса, наши рассуждения приводят к хорошо известной теореме: касательная, проведенная в любой точке гиперболы, делит пополам угол между отрезками, проведенными из фокусов в точку касания.

Может показаться странным, что приходится решать задачу о минимуме, если точки P и Q лежат по одну сторону L, тогда как если точки лежат по разные стороны L, мы рассматриваем задачу о максимуме. Но нетрудно прийти к заключению, что указанное различие совершенно естественно. В первой задаче при удалении по прямой L в бесконечность — в одну или в другую сторону — каждое из расстояний p и q, следовательно, и их сумма, неограниченно возрастает. Таким образом, было бы невозможно найти наибольшее значение $p + q$, и единственной возможной является постановка задачи о *минимуме.* Дело обстоит совершенно иначе во второй задаче, когда P и Q лежат по разные стороны L. В этом случае не будем смешивать три различные величины: разность $p - q$, обратную разность $q - p$ и абсолютную величину $|p - q|$; именно, для последней величины мы определяли *максимум.* Как обстоит дело, легче всего понять, если представить себе, что точка R движется по прямой L, занимая различные положения R_1, R_2, R_3, … Существует такое положение R, для которого разность $p - q$ обращается в нуль; при этом прямая L пересекается с перпендикуляром к отрезку PQ, проведенным из его середины. Ясно, что при этом положении точка R дает минимум для абсолютной величины $|p - q|$. Но по одну сторону от этой точки p больше, чем q, по другую — меньше; значит, величина $p - q$ положительна по одну сторону точки и отрицательна — по другую. Сле-

довательно, сама эта величина не имеет ни максимума, ни минимума в точке, где $|p - q| = 0$. С другой стороны, та точка, в которой $|p - q|$ имеет максимум, наверняка дает экстремум для $p - q$. Если $p > q$, то имеется максимум для $p - q$; если $q > p$, то максимум для $q - p$ и, значит, минимум для $p - q$. Имеется ли максимум или минимум для $p - q$, это зависит от положения двух данных точек относительно прямой L.

В случае, если P и Q находятся на равных расстояниях от L, решения задачи о максимуме, как мы видели, нет вовсе, так как прямая $P'Q$ (см. рис. 182) параллельна L. И тогда при удалении R в бесконечность в том или в другом направлении величина $|p - q|$ стремится к некоторому конечному пределу. Этот предел есть не что иное, как длина s проекции отрезка PQ на прямую L (читатель может доказать это в качестве упражнения). Величина $|p - q|$ при рассматриваемых обстоятельствах всегда меньше, чем предел s, и максимума не существует, так как, какова бы ни была данная точка R, всегда можно указать другую, более удаленную, для которой $|p - q|$ будет больше и, однако, еще не совсем равно s.

***5. Экстремальные расстояния точки от данной кривой.** Начнем с того, что определим наибольшее и наименьшее расстояния данной точки P от точек данной кривой C. Предположим для простоты, что C есть простая замкнутая кривая, имеющая всюду касательную (рис. 184). Понятие касательной к кривой, принимаемое здесь на интуитивной основе, будет подвергнуто анализу в следующей главе. Ответ очень прост: если для некоторой точки R на C расстояние PR достигает минимума или максимума, то прямая PR непременно перпендикулярна к касательной к C в точке R; короче говоря, прямая PR перпендикулярна к C. Доказательство вытекает из следующего обстоятельства: окружность с центром P, проходящая через R, должна быть касательной к кривой C. Действительно, если R есть точка наименьшего расстояния, то кривая C должна целиком лежать вне круга и поэтому в точке R не может его пересекать; если же R есть точка

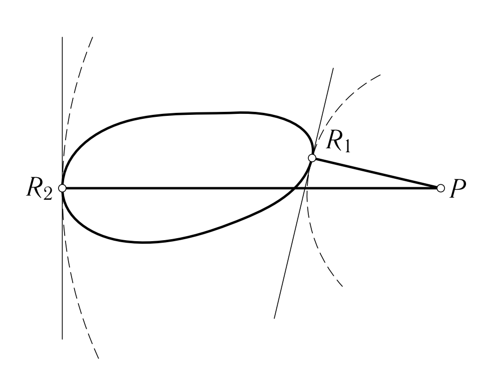


Рис. 184. Экстремальные расстояния до точек кривой

наибольшего расстояния, то C должна целиком лежать внутри круга и потому опять-таки в точке R пересекать его не может. (Это следует из того очевидного факта, что расстояние некоторой точки от P меньше,

чем RP, если эта точка внутри круга, и больше, чем RP, если она вне его.) Итак, окружность и кривая касаются в точке R, и касательная у них в этой точке одна и та же. Остается заметить, что отрезок PR как радиус окружности перпендикулярен к касательной к окружности в точке R и, следовательно, к самой кривой C в той же точке.

В теснейшей связи с предыдущим стоит следующее предложение, доказательство которого предоставляется читателю: диаметр замкнутой кривой C (т. е. наибольшая из ее хорд) в своих концах обязательно перпендикулярен к C. Аналогичное утверждение можно сформулировать и доказать для трехмерного случая.

Упражнение. Докажите, что наикратчайший и наидлиннейший отрезки, связывающие две взаимно непересекающиеся замкнутые кривые, перпендикулярны в своих концах к самым кривым.

Можно обобщить и задачи пункта 4, касающиеся суммы и разности расстояний. Рассмотрим вместо прямой линии L простую замкнутую кривую C, обладающую касательной в каждой точке, и еще две точки P и Q, не лежащие на C. Постараемся охарактеризовать те точки на кривой C, для которых сумма $p+q$ или разность $p-q$ принимают экстремальные значения (причем p и q обозначают соответственно расстояния переменной точки на C от точек P и Q). Теперь уже нельзя применить то простое, основанное на отражении, построение, с помощью которого мы решили обе задачи в случае, когда C была прямой линией. Но мы можем воспользоваться для поставленной здесь цели свойствами эллипса и гиперболы. Так как C на этот раз — замкнутая кривая, а не линия, уходящая в бесконечность, то и максимум и минимум на ней действительно реализуются: в самом деле, можно не подвергать сомнению то обстоятельство, что величины $p+q$ и $p-q$ достигают и наибольшего и наименьшего значений на всяком конечном сегменте кривой, следовательно, на замкнутой кривой (см. § 7).

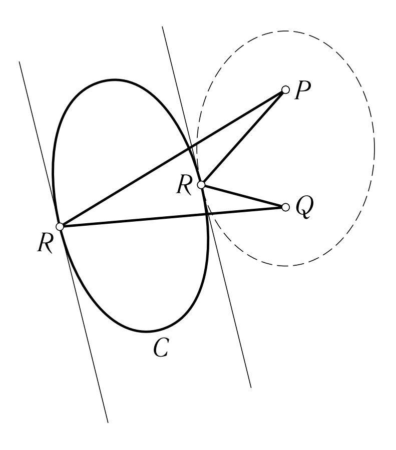


Рис. 185. Максимум и минимум сумм $PR+QR$

Останавливаясь на случае суммы $p+q$, предположим, что R — та точка на C, в которой имеет место максимум; пусть $2a$ есть значение $p+q$ в этой точке. Рассмотрим эллипс с фокусами P и Q — геометрическое место точек, для которых $p+q=2a$. Этот эллипс в точке R должен касаться кривой C (доказательство предоставляется читателю в

качестве упражнения). Но мы видели, что отрезки PR и QR образуют одинаковые углы с эллипсом в точке R, и так как эллипс в точке R касается кривой C, то отрезки PR и QR образуют в той же точке также одинаковые углы и с C. Совершенно аналогичное рассуждение приводит нас к тому же результату и в случае, если в точке R сумма $p+q$ обращается в минимум.

Итак, мы пришли к теореме: *дана замкнутая кривая C и две точки P и Q вне ее; тогда в каждой из точек R, в которых сумма $p+q$ принимает наибольшее или наименьшее значение на кривой C, отрезки PR и QR образуют одинаковые углы с самой кривой* (*т. е. с ее касательной*).

Если точка P внутри C, а точка Q вне C, то теорема остается справедливой для той точки, где $p+q$ принимает наибольшее значение, но она теряет смысл для точки, где $p+q$ принимает наименьшее значение, так как эллипс вырождается в отрезок прямой.

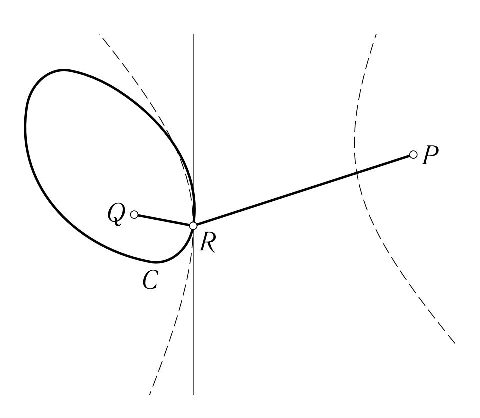


Рис. 186. Минимум разности $PR - QR$

Рассуждая аналогичным образом (воспользовавшись вместо свойств эллипса свойством гиперболы), читатель сможет доказать следующую теорему: *дана замкнутая кривая C и две точки P и Q — одна внутри, другая вне C; тогда в каждой из тех точек R на C, где разность $p-q$ принимает наибольшее или наименьшее значение, отрезки PR и QR образуют одинаковые углы с самой кривой C*. Но нужно вместе с тем отметить, что между случаем, когда C — прямая, и случаем, когда C — замкнутая кривая, есть существенное различие: в первом случае приходится разыскивать максимум абсолютной величины разности, т. е. максимум $|p-q|$, тогда как во втором сама разность $p-q$ достигает и наибольшего и наименьшего значений.

§ 2. Общий принцип, которому подчинены экстремальные задачи

1. Принцип. Предыдущие задачи являются частными случаями некоторой общей проблемы, которую удобнее всего сформулировать аналитически. Возвращаясь к первой из рассмотренных задач, касающейся суммы $p+q$, мы видим, что она заключается в том, чтобы, обозначив через x, y координаты точки R, через x_1, y_1, координаты точки P и через

x_2, y_2 координаты точки Q, найти экстремальные значения функции

$$f(x,\, y) = p + q,$$

где положено

$$p = \sqrt{(x - x_1)^2 + (y - y_1)^2}, \qquad q = \sqrt{(x - x_2)^2 + (y - y_2)^2}.$$

Рассматриваемая функция непрерывна во всей плоскости, но точка R с координатами x, y подчинена требованию находиться на кривой C. Эта последняя кривая, допустим, определена уравнением $g(x,\, y) = 0$; например, уравнением $x^2 + y^2 - 1 = 0$, если C — единичная окружность.

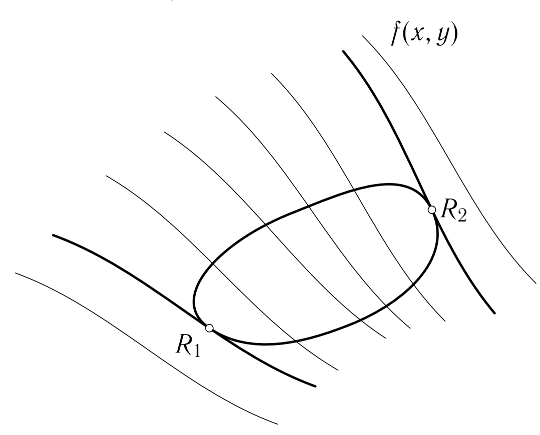


Рис. 187. Экстремумы функции на кривой

Обратимся теперь к общей задаче: найти экстремальные значения некоторой данной функции $f(x,\, y)$, если переменные x и y подчинены условию $g(x,\, y) = 0$. Постараемся охарактеризовать решение этой задачи. Для этого рассмотрим семейство кривых $f(x,\, y) = c$; при этом под «семейством» кривых понимаем совокупность всех кривых, определяемых указанным уравнением при различных значениях постоянной c (но такое значение неизменно для всех точек каждой кривой в отдельности). Предположим, что через каждую точку плоскости — или по крайней мере некоторой ее части, содержащей кривую C, — проходит одна и только одна кривая семейства $f(x,\, y) = c$. Тогда при непрерывном увеличении c кривая $f(x,\, y) = c$ «заметает» некоторую часть плоскости, однако при этом ни одну точку не «заметает» дважды. (Примеры такого рода семейств: $x^2 + y^2 = c$, $x + y = c$, $x = c$.) В частности, одна кривая рассматриваемого семейства пройдет через точку R_1, в которой $f(x,\, y)$ принимает наибольшее значение на кривой C, и другая — через точку R_2, в которой $f(x,\, y)$ принимает наименьшее значение на C. Пусть наибольшее значение есть a, наименьшее — b. По одну сторону кривой $f(x,\, y) = a$ значение $f(x,\, y) = a$ меньше, чем a, по другую — больше, чем a. Так как

на кривой C имеет место неравенство $f(x, y) \leqslant a$, то кривая C должна целиком лежать по одну и ту же сторону кривой $f(x, y) = a$; отсюда следует, что она в точке R_1 касается кривой $f(x, y) = a$. Точно так же кривая C касается в точке R_2 кривой $f(x, y) = b$.

Итак, доказана общая теорема: *если в точке R на кривой C функция $f(x, y)$ имеет экстремальное значение a, то кривая $f(x, y) = a$ в точке R касается кривой C.*

2. Примеры. Легко понять, что ранее полученные результаты являются частным случаем этой общей теоремы. Если речь идет об экстремуме суммы $p + q$, то функция $f(x, y)$ есть $p + q$, а кривые $f(x, y) = c$ — софокусные эллипсы с фокусами P и Q. В согласии с общей теоремой эллипсы, проходящие через те точки кривой C, где достигается экстремум одного или другого вида, касаются кривой C в этих точках. Если речь идет об экстремуме разности $p - q$, то функция $f(x, y)$ есть $p - q$, и тогда кривые $f(x, y) = c$ — софокусные гиперболы с фокусами P и Q; и в этом случае гиперболы, проходящие через точки, где достигается экстремум, касаются кривой C.

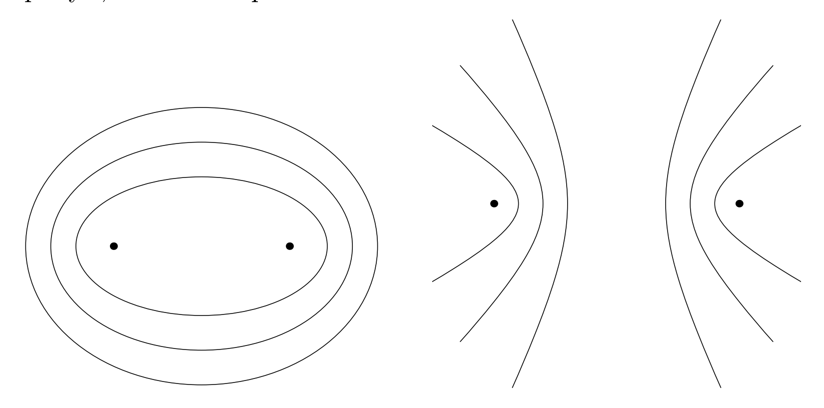


Рис. 188. Софокусные эллипсы Рис. 189. Софокусные гиперболы

Вот еще пример задачи того же типа. Дан отрезок прямой PQ и прямая l, его не пересекающая; требуется установить: из какой точки l отрезок PQ виден под наибольшим углом?

Функция, максимум которой надлежит определить в этой задаче, есть угол θ, под которым из точки, движущейся по прямой l, виден отрезок PQ; если R — какая угодно точка плоскости с координатами x, y, то угол, под которым отрезок PQ виден из R, есть функция $\theta = f(x, y)$ от переменных x, y. Из элементарной геометрии известно, что семейство кривых $\theta = f(x, y) = \text{const}$ состоит из окружностей, проходящих через P и Q, так как хорда круга видна под одним и тем же углом из всех точек дуги окружности, расположенной по одну сторону хорды. Из рис. 190

видно, что, вообще говоря, имеется две окружности рассматриваемого семейства, касающиеся прямой l: их центры расположены по разные стороны отрезка PQ. Одна из точек касания дает абсолютный максимум величины θ, тогда как другая — лишь «относительный» максимум: это значит, что значения θ в этой точке больше, чем значения в некоторой окрестности рассматриваемой точки. Больший из двух максимумов — абсолютный максимум — дается той точкой касания, которая расположена в остром угле, образованном прямой l и продолжением отрезка PQ, а меньший — той точкой касания, которая расположена в тупом угле, образованном этими прямыми. (Точка пересечения прямой l с продолжением отрезка PQ дает минимальное значение угла θ, именно $\theta = 0$.)

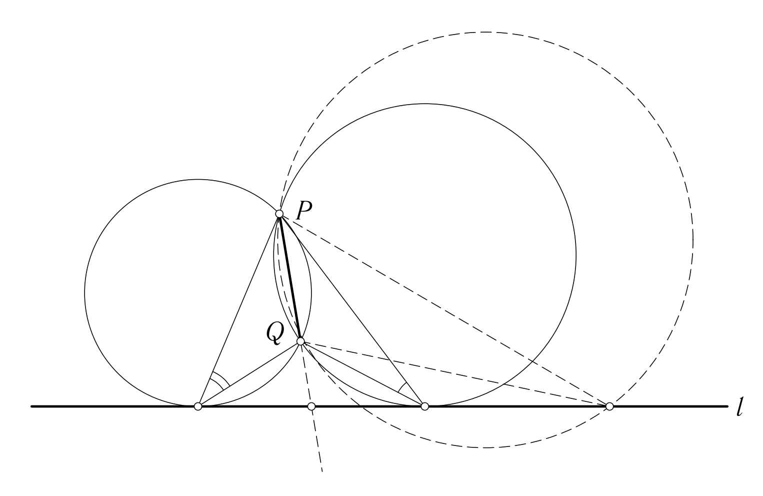


Рис. 190. Из какой точки l отрезок PQ виден под наибольшим углом?

Обобщая рассмотренную задачу, мы можем заменить прямую l какой-нибудь кривой C и искать точки R на кривой C, из которых данный отрезок PQ, не пересекающий C, виден под наибольшим или наименьшим углом. В этой задаче, как и в предыдущей, окружность, проходящая через P, Q и R, должна в точке R касаться кривой C.

§ 3. Стационарные точки и дифференциальное исчисление

1. Экстремальные и стационарные точки. В предшествующих рассуждениях мы совсем не пользовались техническими приемами дифференциального исчисления.

Трудно не признать, что наши элементарные методы являются более простыми и более прямыми, чем методы анализа. Вообще, занимаясь той или иной научной проблемой, лучше исходить из ее индивидуальных

особенностей, чем полагаться исключительно на общие методы, хотя, с другой стороны, общий принцип, уясняющий смысл применяемых специальных процедур, конечно, всегда должен играть руководящую роль. Таково именно значение методов дифференциального исчисления при рассмотрении экстремальных проблем. Наблюдаемое в современной науке стремление к общности представляет только одну сторону дела, так как то, что в математике является подлинно жизненным, без всякого сомнения обусловливается индивидуальными чертами рассматриваемых проблем и применяемых методов.

В своем историческом развитии дифференциальное исчисление в весьма значительной степени испытало воздействие индивидуальных проблем, связанных с разысканием наибольших и наименьших значений величин. Связь между экстремальными проблемами и дифференциальным исчислением можно уяснить себе следующим образом. В главе VIII мы займемся обстоятельным изучением производной $f'(x)$ от функции $f(x)$ и ее геометрического смысла. Там мы увидим, что, говоря кратко, производная $f'(x)$ есть наклон касательной к кривой $y = f(x)$ в точке (x, y). Геометрически очевидно, что в точках максимума или минимума гладкой кривой $y = f(x)$ касательная к кривой непременно должна быть горизонтальной, т. е. наклон должен равняться нулю. Таким образом, мы получаем для точек экстремума условие $f'(x) = 0$.

Чтобы отдать себе ясно отчет в том, что означает обращение в нуль производной $f'(x)$, рассмотрим кривую, изображенную на рис. 191. Мы видим здесь пять точек A, B, C, D, E, в которых касательная к кривой горизонтальна; обозначим соответствующие значения $f(x)$ в этих точках через a, b, c, d, e. Наибольшее значение $f(x)$ (в пределах области, изображенной на чертеже) достигается в точке D, наименьшее — в точке A. В точке B имеется максимум — в том смысле, что во всех точках *некоторой окрестности* точки B значение $f(x)$ меньше, чем b, хотя в точках, близких к D, значение $f(x)$ все же больше, чем b. По этой причине принято говорить, что в точке B имеется *относительный максимум* функции $f(x)$, тогда как в точке D — *абсолютный максимум.* Точно так же в точке C имеет место *относительный минимум*, а в точке A — *абсолютный минимум.* Наконец, что касается точки E, то в ней нет ни максимума, ни минимума, хотя в ней все же осуществляется равенство $f'(x) = 0$. Отсюда следует, что обращение в нуль производной $f'(x)$ есть *необходимое*, но никак не *достаточное* условие для появления экстремума гладкой функции $f(x)$; другими словами, во всякой точке, где имеется экстремум (абсолютный или относительный), непременно имеет место равенство $f'(x) = 0$, но не во всякой точке, где $f'(x) = 0$, обязан быть экстремум. Те точки, в которых производная $f'(x)$ обращается в нуль, — независимо от того, имеется ли в них экстремум, — называются *стационарными.* Дальнейший анализ приводит к более или менее

сложным условиям, касающимся высших производных функции $f(x)$ и полностью характеризующим максимумы, минимумы и иные стационарные точки.

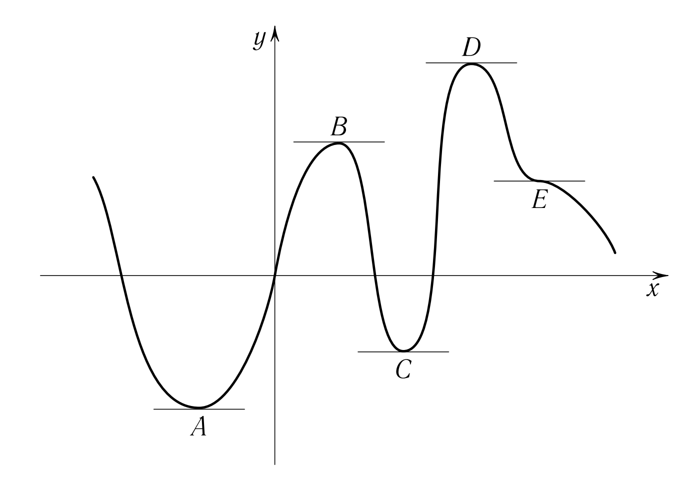


Рис. 191. Стационарные точки функции

2. Максимумы и минимумы функций нескольких переменных. Седловые точки. Существуют экстремальные проблемы, которые не могут быть выражены с помощью понятия функции $f(x)$ от одной переменной. Простейшим относящимся сюда примером является проблема нахождения экстремумов функции $z = f(x, y)$ от двух независимых переменных.

Мы всегда можем представлять себе функцию $f(x, y)$ как высоту z поверхности над плоскостью x, y, и эту картину будем интерпретировать, скажем, как горный ландшафт. Максимум функции $f(x, y)$ соответствует горной вершине, минимум — дну ямы или озера. В обоих случаях, если только поверхность гладкая, касательная плоскость к поверхности обязательно горизонтальна. Но, помимо вершин гор и самых низких точек в ямах, могут существовать и иные точки, в которых касательная плоскость горизонтальна: это «седловые» точки, соответствующие горным перевалам. Исследуем их более внимательно. Предположим (рис. 192), что имеются две вершины A и B в горном хребте и две точки C и D на различных склонах хребта; предположим, что из C нам нужно пройти в D. Рассмотрим сначала те пути, ведущие из C в D, которые получаются при пересечении поверхности плоскостями, проходящими через C и D. Каждый такой путь имеет самую высокую точку. При изменении положения секущей плоскости меняется и путь, и можно будет найти такой путь, для которого *наивысшая* точка будет в

самом низком из возможных положений. Наивысшей точкой E на этом пути является точка горного перевала в нашем ландшафте; ее можно назвать также *седловой точкой*. Ясно, что в точке E нет ни максимума, ни минимума, так как сколь угодно близко к E существуют на поверхности такие точки, которые выше E, и такие, которые ниже E. Можно было бы в предыдущем рассуждении и не ограничиваться рассмотрением только тех путей, которые возникают при пересечении поверхности плоскостями, а рассматривать какие угодно пути, соединяющие C и D. Характеристика, данная нами точке E, от этого бы не изменилась.

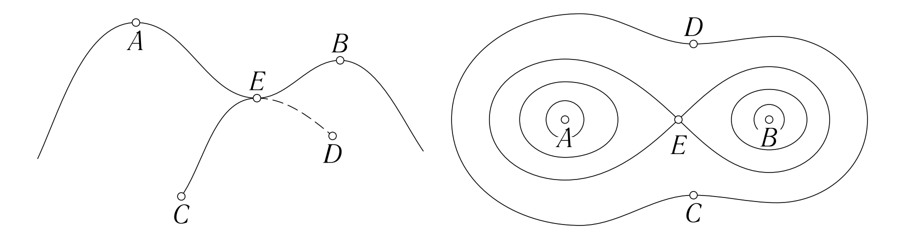


Рис. 192. Горный перевал

Рис. 193. Соответствующая карта с линиями уровня

Точно так же, если бы мы пожелали от вершины A пройти к вершине B, то всякий путь, который мы могли бы выбрать, имел бы самую низкую точку; рассматривая хотя бы только плоские сечения, мы нашли бы такой путь AB, для которого наименьшая точка была бы расположена наиболее высоко, причем получилась бы опять прежняя точка E. Таким образом, эта седловая точка E обладает свойством доставлять самый высокий минимум или самый низкий максимум: здесь имеет место «максиминимум» или «минимаксимум» — сокращенно *минимакс*. Касательная плоскость в точке E горизонтальна; действительно, так как E — наинизшая точка пути AB, то касательная к AB в E горизонтальна, и аналогично, так как E — наивысшая точка пути CD, то и касательная к CD в E горизонтальна. Поэтому касательная плоскость, обязательно проходящая через эти две касательные прямые, горизонтальна. Итак, мы обнаруживаем три различных типа точек с горизонтальными касательными плоскостями: точки максимума, точки минимума и, наконец, седловые точки; соответственно существует и три различных типа стационарных значений функции.

Другой способ представлять геометрически функцию $f(x,\, y)$ заключается в вычерчивании линий уровня — тех самых, которые употребляются в картографии для обозначения высот на местности (см. стр. 308). Линией уровня называется такая кривая в плоскости x, y, вдоль которой функция $f(x,\, y)$ имеет одно и то же значение; другими словами, линии уровня — то же, что и кривые семейства $f(x,\, y) = c$. Через обыкновенную

точку плоскости проходит в точности одна линия уровня; точки максимума и минимума бывают окружены замкнутыми линиями уровня, в седловых точках пересекаются две (или более) линии уровня. На рис. 193 проведены линии уровня, соответствующие ландшафту, изображенному на рис. 192.

При этом особенно наглядным становится замечательное свойство седловой точки E: всякий путь, связывающий A и B и не проходящий через E, частично лежит в области, где $f(x, y) < f(E)$, тогда как путь AEB на рис. 192 имеет минимум как раз в точке E. Таким же образом мы убеждаемся, что значение $f(x, y)$ в точке E представляет собой наименьший максимум на путях, связывающих C и D.

3. Точки минимакса и топология. Существует глубокая связь между общей теорией стационарных точек и топологическими идеями. По этому поводу мы можем здесь дать только краткое указание и ограничимся рассмотрением одного примера.

Рассмотрим горный ландшафт на кольцеобразном острове B с двумя береговыми контурами C и C'; если обозначим, как раньше, высоту над уровнем моря через $u = f(x, y)$, причем допустим, что $f(x, y) = 0$ на контурах C и C' и $f(x, y) > 0$ внутри, то на острове должен существовать по меньшей мере один горный перевал: на рис. 194 такой перевал находится в точке, где пересекаются две линии уровня. Справедливость высказанного утверждения становится наглядной, если мы поставим своей задачей найти такой путь, соединяющий C и C', который не поднимался бы на бóльшую высоту, чем это неизбежно. Каждый путь от C к C' имеет наивысшую точку, и если мы выберем такой путь, для которого наивысшая точка оказывается самой низкой, то полученная таким образом наивысшая точка и будет седловой точкой функции $u = f(x, y)$. (Следует оговорить представляющий исключение тривиальный случай, когда некоторая горизонтальная плоскость касается кольцеобразного горного хребта по замкнутой кривой.) В случае области, ограниченной p замкнутыми кривыми, вообще говоря, должно существовать не менее чем $p - 1$ точек минимакса. Подобного же рода соотношения, как установил Марстон Морс, имеют место и для многомерных областей,

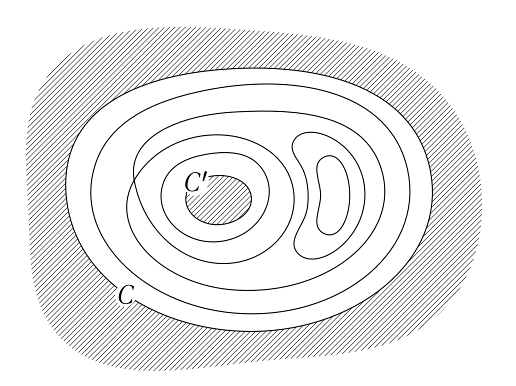


Рис. 194. Стационарные точки в двусвязной области

но разнообразие топологических возможностей и типов стационарных точек в этом случае значительно большее. Эти соотношения образуют основу современной теории стационарных точек.

4. Расстояние точки от поверхности. Для расстояний точки P от различных точек замкнутой кривой существуют (по меньшей мере) два стационарных значения: минимальное и максимальное. При переходе к трем измерениям не обнаруживается никаких новых фактов, если мы ограничимся рассмотрением такой поверхности C, которая топологически эквивалентна сфере (как, например, эллипсоид). Но если поверхность рода 1 или более высокого, то дело обстоит иначе. Рассмотрим поверхность тора C. Какова бы ни была точка P, всегда, конечно, существуют на торе C точки, дающие наибольшее и наименьшее расстояние от P, причем соответствующие отрезки перпендикулярны к самой поверхности. Но мы сейчас установим, что в этом случае существуют и точки минимакса. Вообразим на торе один из «меридианных» кругов L (рис. 195) и на этом круге L найдем точку Q, ближайшую к P. Затем, перемещая круг L по тору, найдем такое его положение, чтобы расстояние PQ стало: а) минимальным — тогда получается точка на C, ближайшая к P; б) максимальным — тогда получится стационарная точка минимакса. Таким же образом мы могли бы найти на L точку, наиболее удаленную от P, и затем искать положение L, при котором найденное наибольшее расстояние было бы: в) максимальным (получится точка на C, наиболее удаленная от P), г) минимальным. Итак, мы получим четыре различных стационарных значения для расстояния точки тора C от точки P.

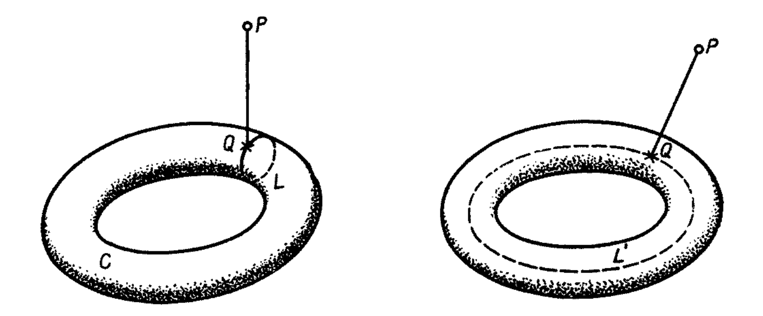


Рис. 195–196. Расстояние от точки до поверхности

Упражнение. Повторите то же рассуждение для иного типа L' замкнутой кривой на C, которая также не может быть стянута в точку (рис. 196).

§ 4. Треугольник Шварца

1. Доказательство, предложенное Шварцем. Герман Амандус Шварц (1843–1921), выдающийся математик, профессор Берлинского университета, сделал многое для развития современной теории функций и анализа. Он не считал ниже своего достоинства писать на темы элементарного содержания, и одна из его работ посвящена следующей задаче: в данный остроугольный треугольник вписать другой треугольник с минимальным периметром. (Говоря, что некоторый треугольник вписан в данный, мы подразумеваем, что на каждой из сторон данного треугольника имеется вершина рассматриваемого треугольника.) Мы убедимся в дальнейшем, что существует только один искомый треугольник: именно, его вершинами являются основания высот данного треугольника. Такой треугольник условимся называть *высотным* треугольником.

Шварц доказал минимальное свойство высотного треугольника, применяя метод отражения и основываясь на следующей теореме элементарной геометрии: *в каждой из вершин P, Q, R* (рис. 197) *две стороны высотного треугольника образуют одинаковые углы со стороной данного треугольника, именно, каждый из этих углов равен углу при противоположной вершине данного треугольника.* Например, углы ARQ и BRP равны каждый углу C и т. д.

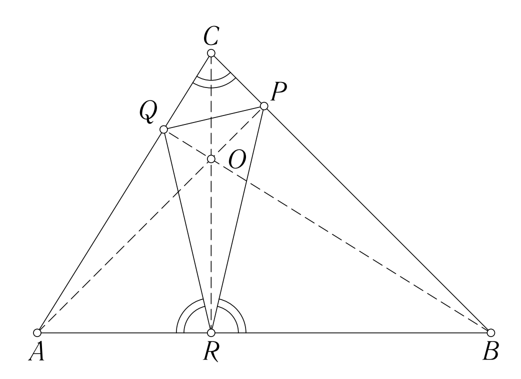


Рис. 197. Высотный треугольник в треугольнике ABC

Докажем прежде всего эту теорему. Так как углы OPB и ORB прямые, то около четырехугольника $OPBR$ можно описать окружность. Следовательно, $\angle PBO = \angle PRO$, так как названные углы опираются на одну и ту же дугу описанной окружности. Но угол PBO дополнительный к углу C, так как треугольник CBQ прямоугольный, а угол PRO дополнительный к углу PRB. Поэтому $\angle PRB = \angle C$. Таким же образом, рассуждая по поводу четырехугольника $QORA$, заключаем, что $\angle QRA = \angle C$ и т. д.

Этот результат приводит к следствию, относящемуся к высотному треугольнику: так как, например, $\angle AQR = \angle CQP$, то при отражении относительно стороны AC данного треугольника сторона RQ направляется по стороне PQ, и обратно. Аналогично для других сторон.

Перейдем теперь к доказательству минимального свойства высотного треугольника. В треугольнике ABC рассмотрим, наряду с высотным треугольником, какой-нибудь другой вписанный треугольник, скажем, UVW. Отразим всю фигуру сначала относительно стороны AC треугольника ABC, затем вновь получившуюся фигуру отразим относительно стороны AB, потом — относительно BC, потом — относительно AC и, наконец, относительно AB. Таким образом мы получим всего шесть конгруэнтных треугольников, причем в каждом из них будет заключен высотный треугольник и еще другой вписанный треугольник (рис. 198). Сторона BC последнего треугольника параллельна стороне BC первого треугольника. В самом деле, при первом отражении сторона BC поворачивается по часовой стрелке на угол $2C$, затем опять по часовой стрелке на угол $2B$; при третьем отражении — остается неизменной; при четвертом — поворачивается на угол $2C$ против часовой стрелки и при пятом — на угол $2B$ опять против часовой стрелки. Итого, общий угол поворота равен нулю.

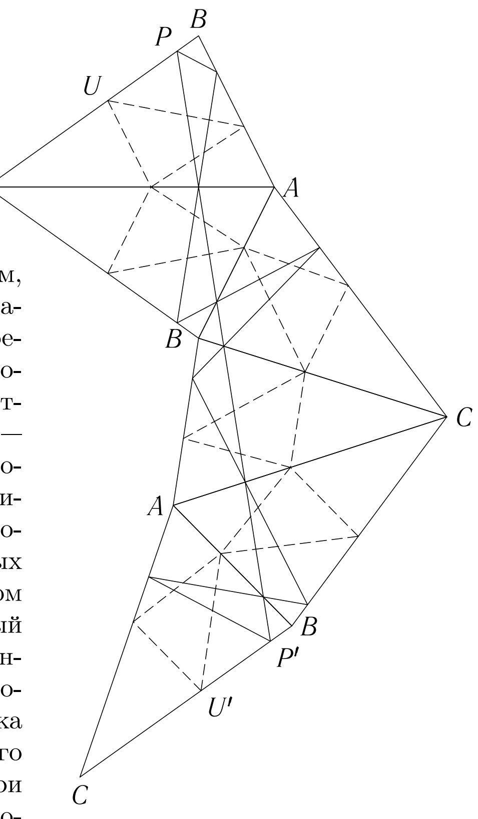


Рис. 198. Доказательство минимального свойства высотного треугольника, данное Шварцем

Благодаря указанному выше свойству высотного треугольника прямолинейный отрезок PP' равен удвоенному периметру треугольника PQR: действительно, PP' составляется из шести отрезков, по очереди равных первой, второй и третьей стороне PQR, причем каждая из

сторон входит дважды. Таким же образом ломаная линия, соединяющая U и U', имеет длину, равную удвоенному периметру треугольника UVW. Эта ломаная не короче, чем прямолинейный отрезок UU'. Что же касается прямолинейного отрезка UU', то он равен PP', так как отрезок UU' параллелен PP'. Значит, ломаная линия UU' не короче, чем прямая PP', т. е. периметр высотного треугольника не больше, чем периметр любого другого треугольника, вписанного в данный. Это и нужно было доказать. Итак, установлено, что минимум существует и что он реализуется в случае высотного треугольника. Что нет иного вписанного треугольника с тем же периметром — это, однако, не доказано, и это мы докажем дальше.

2. Другое доказательство. Следующее решение задачи Шварца является, вероятно, самым простым. Оно основывается на теореме, ранее доказанной в этой главе: если точки P и Q лежат по одну сторону прямой L (но не на ней самой), то сумма расстояний $PR + RQ$, где R — точка на L, обращается в минимум в том случае, если PR и QR образуют одинаковые углы с L. Пусть треугольник PQR, вписанный в данный треугольник ABC, решает поставленную минимальную задачу. Тогда точка R на стороне AB должна быть такой, чтобы сумма $PR + QR$ была наименьшей, следовательно, углы ARQ и BRP должны быть равны; и точно так же $\angle AQR = \angle CQP$, $\angle BPR = \angle CPQ$. Таким образом, для искомого треугольника с минимальным периметром — если только таковой существует — должно быть выполнено то же самое свойство равенства углов, каким обладает высотный треугольник. Остается показать, что при таком условии наш треугольник не может отличаться от высотного. Кроме того, так как в теореме, на которую мы ссылались,

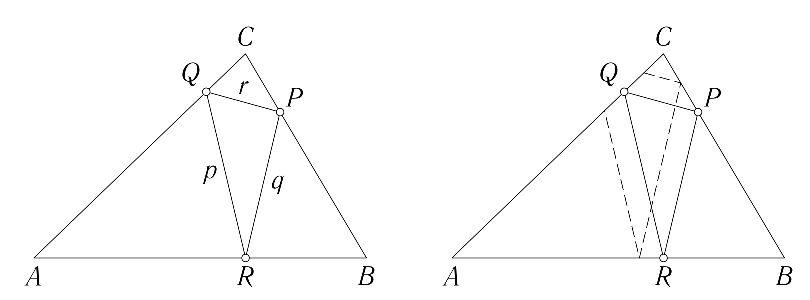


Рис. 199–200. Другое доказательство минимального свойства высотного треугольника

предполагается, что P и Q не лежат на AB, то доказательство не годится для случая, когда одна из точек P, Q, R совпадает с какой-нибудь вершиной данного треугольника (при этом периметр треугольника выродился бы в удвоенную соответствующую высоту); чтобы доказательство

было исчерпывающим, нужно еще установить, что периметр высотного треугольника меньше любой из удвоенных высот данного треугольника.

Обращаясь сначала к первому пункту, заметим, что если вписанный треугольник обладает указанным выше свойством равенства углов, то рассматриваемые углы при вершинах P, Q и R соответственно равны углам A, B и C. В самом деле, допустим, например, что $\angle ARQ = \angle BRP = \angle C + \delta$. Тогда, применяя теорему о сумме углов треугольника к треугольникам ARQ и BRP, мы видим, что углы при Q должны равняться $B - \delta$, а углы при P должны равняться $A - \delta$. Но тогда сумма углов треугольника CPQ равна $(A - \delta) + (B - \delta) + C = 180^\circ - 2\delta$; с другой стороны, она же равна 180°. Поэтому $\delta = 0$. Мы уже видели, что высотный треугольник обладает отмеченным свойством. Всякий иной вписанный треугольник, обладающий тем же свойством, имел бы стороны, соответственно параллельные сторонам высотного треугольника; другими словами, он был бы ему подобен и подобно расположен. Читатель докажет самостоятельно, что, кроме самого высотного треугольника, другого такого треугольника не существует (рис. 200).

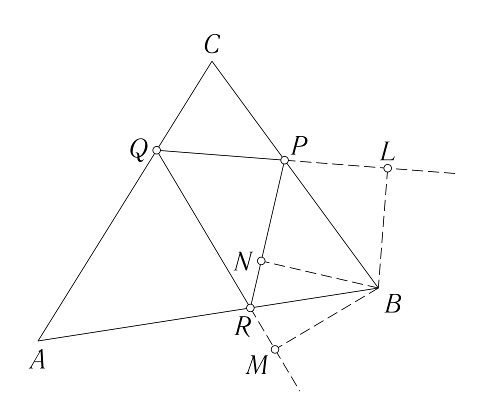


Рис. 201. К доказательству минимального свойства высотного треугольника

Покажем, наконец, по-прежнему ограничиваясь случаем остроугольного треугольника, что периметр высотного треугольника меньше, чем любая удвоенная высота данного треугольника. Проведем прямые QP и QR и затем из вершины B (рис. 201) опустим перпендикуляры на прямые QP, QR и PR; пусть L, M и N — основания этих перпендикуляров. Так как отрезки QL и QM являются проекциями высоты QB на прямые QP и QR, то $QL + QM < 2QB$. Но $QL + QM = p$, где через p обозначен периметр высотного треугольника. Действительно, треугольники MRB и NRB равны, так как $\angle MRB = \angle NRB$, а углы при вершинах M и N прямые. Значит, $RM = RN$; и поэтому $QM = QR + RN$. Точно так же мы убеждаемся, что $PN = PL$ и, следовательно, $QL = QP + PN$. Отсюда вытекает: $QL + QM = QP + PN + QR + RN = QP + PR + RQ = p$. Но раньше было показано, что $2QB > QL + QM$. Итак, p меньше, чем удвоенная высота QB. Это же рассуждение может быть применено и к каждой из двух других высот. Таким образом, минимальное свойство высотного треугольника доказано полностью.

Между прочим, приведенное построение позволяет непосредственно вычислить p. Мы знаем, что углы PQC и RQA равны углу B, так что $\angle PQB = \angle RQB = 90^\circ - B$ и $\cos\angle PQB = \sin B$. Отсюда следует, с помощью элементарных тригонометрических соображений, что $QM = QL = QB\sin B$, и $p = 2QB\sin B$. Таким же образом можно показать, что $p = 2PA\sin A = 2RC\sin C$. Из тригонометрии известно, что $RC = a\sin B = b\sin A$ и т. д., откуда следует: $p = 2a\sin B\sin C = 2b\sin C\sin A = 2c\sin A\sin B$. И наконец, вводя радиус описанного круга r и принимая во внимание, что $a = 2r\sin A$, $b = 2r\sin B$, $c = 2r\sin C$, мы получим симметрическую формулу

$$p = 4r\sin A\sin B\sin C.$$

3. Тупоугольные треугольники. В обоих предшествующих доказательствах предполагалось, что все три угла A, B, C острые. Если бы, скажем, угол C был тупой (рис. 202), то точки P и Q лежали бы вне треугольника. Поэтому, строго говоря, высотный треугольник уже нельзя было бы считать вписанным в данный, если только мы не условимся заранее называть вписанным такой треугольник, вершины которого лежат на сторонах данного треугольника или на их продолжениях. Как бы то ни было, высотный треугольник в расширенном смысле не обладает минимальным периметром, так как $PR > CR$, $QR > CR$, и, значит, $p = PR + QR + PQ > > 2CR$. Так как рассуждение в первой части последнего доказательства показывает, что минимальный периметр — если только он не дается высотным треугольником — должен быть равен одной из удвоенных высот, то отсюда легко заключить, что в случае тупоугольного треугольника «вписанный треугольник» с наименьшим периметром есть не что иное, как высота, опущенная из вершины тупого угла, учитываемая в обоих направлениях; хотя треугольника в собственном смысле здесь и нет, однако можно все же указать настоящие вписанные треугольники, периметры которых как угодно мало отличаются от удвоенной высоты. В промежуточном случае, когда данный треугольник прямоугольный, оба решения (высотный треугольник и удвоенная высота, опущенная из прямого угла) совпадают.

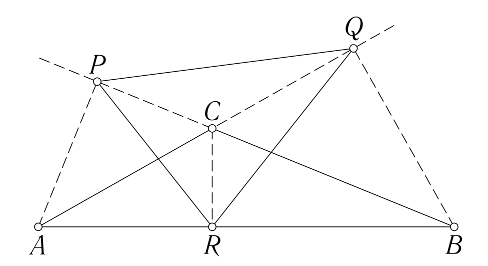


Рис. 202. Высотный треугольник в тупоугольном треугольнике

Не лишен интереса вопрос о том, не обладают ли каким-нибудь экстремальным свойством высотные треугольники данных тупоугольных треугольников. Не имея возможности подробно рассматривать этот вопрос, отметим лишь, что такие высотные треугольники не обращают в минимум сумму сторон $p + q + r$, но зато обеспечивают стационар-

ное значение типа минимакса для выражения вида $p+q-r$, где r — та сторона вписанного (в расширенном смысле) треугольника, которая соответствует тупому углу.

4. Треугольники, образованные световыми лучами. Если допустим, что треугольник ABC изображает комнату с зеркальными стенами, то высотный треугольник определяет единственный треугольный контур, который может быть образован световым лучом. Другие замкнутые многоугольные контуры также не исключены, как показывает рис. 203, но высотный треугольник имеет три стороны.

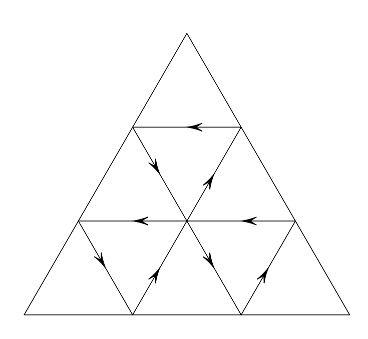


Рис. 203. Замкнутый световой путь в треугольном зеркале

Обобщим рассматриваемую проблему и спросим себя о возможных «световых треугольниках» в произвольной области, ограниченной одной или несколькими гладкими кривыми; точнее говоря, нас интересуют треугольники, вершины которых лежат на заданных кривых, а каждые две прилежащие стороны образуют равные углы с соответствующей кривой. Мы видели в § 1, что равенство углов является необходимым условием как для максимума, так и для минимума суммы соответствующих сторон, так что, смотря по обстоятельствам, могут возникать различные типы световых треугольников. Так, рассматривая внутренность единственной замкнутой гладкой кривой C, мы можем сказать, что вписанный треугольник максимального периметра должен быть «световым треугольником», обладающим вышеописанными свойствами. Или предположим еще, что каждая из вершин треугольника ABC имеет право находиться на ей соответствующей одной из трех замкнутых гладких кривых (идея Марстона Морса). Тогда световые треугольники характеризуются тем свойством, что их периметры имеют стационарные значения. Но такого рода значение может быть минимальным по отношению ко всем трем вершинам A, B, C; или может быть минимальным по отношению к двум каким-либо вершинам и максимальным по отношению к третьей, или минимальным по отношению к одной какой-нибудь из трех и максимальным относительно двух других; или, наконец, максимальным относительно всех трех. Всего, таким образом, существует по меньшей мере $2^3 = 8$ типов световых треугольников, так как по отношению к каждой из вершин, и притом независимо, возможен максимум или минимум.

***5. Замечания, касающиеся задач на отражение и эргодическое движение.** В динамике и в оптике представляется задачей первостепенной важности дать описание пути или «траектории» части-

цы или светового луча в пространстве на протяжении неограниченного промежутка времени. Предполагая, что то или иное приспособление физически принуждает частицу или луч оставаться в некоторой ограниченной части пространства, особенно интересно установить, заполняет ли траектория в пределе эту часть пространства повсюду с приблизительно одинаковой «плотностью». Траектория, обладающая таким свойством, называется *эргодической*. Допущение существования эргодической траектории является исходной гипотезой для применения статистических методов в современных динамических и атомных теориях. Но известно лишь очень немного ситуаций, при которых может быть проведено строгое математическое доказательство «эргодической гипотезы».

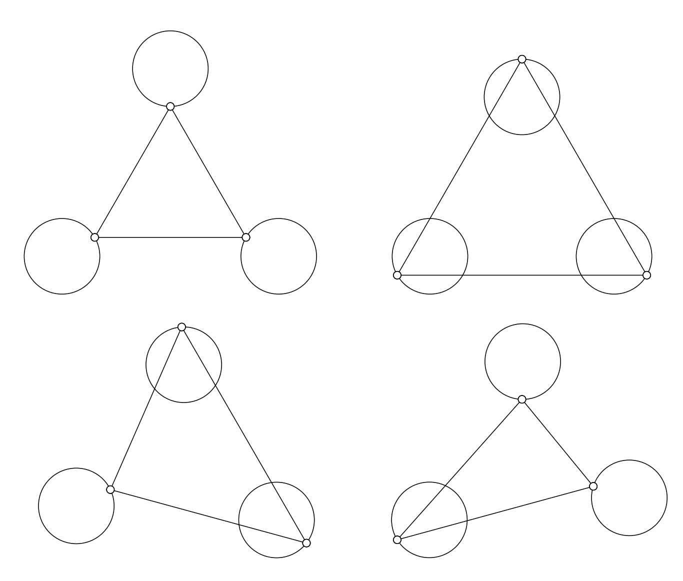


Рис. 204–207. Четыре типа световых треугольников между тремя кругами

Простейшие примеры относятся к случаю, когда движение происходит на плоскости внутри замкнутой кривой C, причем предполагается, что «стенка» C представляет собой математически совершенное зеркало, отражающее частицу (в остальном — свободную) под тем же углом, под каким она падает на стенку. Так, например, прямоугольный ящик — идеализированный бильярдный стол с совершенным отражением, причем рассматриваемая частица играет роль бильярдного шара, — обеспечивает, вообще говоря, эргодическое движение: идеальный «бильярдный шар» на протяжении бесконечного промежутка времени побывает в окрестности любой наперед заданной точки, если только исключить некоторые особые начальные положения и направления движения. Мы не приводим здесь доказательства, впрочем, не представляющего труд-

ностей принципиального порядка.

Особенно любопытно движение на эллиптическом столе с фокусами F_1 и F_2. Так как касательная к эллипсу образует одинаковые углы с отрезками, проведенными из фокусов в точку касания, то каждая траектория, проходящая через один из фокусов, дает отражение, проходящее через другой фокус, и т. д. Нетрудно усмотреть, что после n отражений, независимо от начального положения, траектория при n, неограниченно возрастающем, будет приближаться к большой оси F_1F_2. Если начальный луч не проходит через фокус, то возникают две возможности. Или начальный луч проходит *между фокусами*: тогда все отраженные траектории будут проходить между фокусами, причем будут касательными к некоторой гиперболе с теми же фокусами F_1 и F_2. Или же начальный луч *не разделяет фокусов*: тогда этим же свойством будут обладать все отраженные лучи, причем все они будут касаться некоторого эллипса с теми же фокусами F_1 и F_2. Таким образом, движение внутри эллипса ни при каких начальных условиях не оказывается эргодическим.

Упражнения. 1) Докажите, что если начальный луч проходит через какой-нибудь фокус эллипса, то его n-е отражение при неограниченном возрастании n стремится к большой оси.

2) Докажите, что если начальный луч проходит между фокусами эллипса, то этим же свойством обладают все отраженные лучи, и все они касательны к некоторой гиперболе с фокусами F_1 и F_2; точно так же, если начальный луч не проходит между фокусами, то этим же свойством обладают все отраженные лучи, и все они касательны к некоторому эллипсу с фокусами F_1 и F_2. (*Указание*: установите, что до отражения и после отражения в точке R луч образует соответственно одинаковые углы с отрезками RF_1 и RF_2, потом докажите, что софокусные конические сечения характеризуются отмеченным обстоятельством.)

§ 5. Проблема Штейнера

1. Проблема и ее решение. Очень простая и вместе с тем поучительная проблема была изучена в начале прошлого столетия знаменитым берлинским геометром Якобом Штейнером. Требуется соединить три деревни A, B, C системой дорог таким образом, чтобы их общая протяженность была минимальной. В более точной математической формулировке: на плоскости даны три точки A, B, C; требуется найти такую четвертую точку P, чтобы сумма $a + b + c$ (где a, b, c — расстояния P соответственно от A, B, C) обратилась в минимум. Решение проблемы таково: если в треугольнике ABC все углы меньше $120°$, то в качестве точки P следует взять ту, из которой все три стороны AB, BC, CA видны под углом в $120°$; если же один из углов треугольника ABC, например C, больше или равен $120°$, то точку P нужно совместить с вершиной C.

Обосновать этот результат не представляет труда, если воспользоваться решением уже рассмотренных экстремальных задач. Предположим, что P есть искомая точка. Возможны две альтернативы: или точка P совпадает с одной из вершин A, B, C, или P отлична от всех трех вершин. В первом случае очевидно, что P должна быть вершиной именно самого большого угла C в треугольнике ABC, так как сумма $CA + CB$ меньше, чем сумма каких-нибудь двух других сторон треугольника ABC. Чтобы исчерпать вопрос, остается проанализировать второй возможный случай. Пусть K — окружность с центром C и радиусом c. Тогда точка P должна быть расположена на K таким образом, что $PA + PB$ обращается в минимум. Если обе точки A и B вне K (как на рис. 209), то на основании § 1 отрезки PA и PB должны образовывать одинаковые углы с окружностью K и, следовательно, с радиусом PC, который перпендикулярен к K. Так как это рассуждение можно повторить относительно окружности с центром A и радиусом a, то отсюда следует, что все углы, образованные отрезками PA, PB, PC, равны между собой и, значит, каждый из них равен $120°$, как и было сказано выше. Наше доказательство было построено на допущении, что обе точки A и B находятся вне круга K; докажем, что иначе быть не может. Пусть хотя бы одна из точек A, B, например A, находится внутри окружности K или на ней самой. Тогда $AC \leqslant c$; так как, с другой стороны, при любом расположении точек A, B, P сумма $a + b \geqslant AB$, то $a + b + c \geqslant AB + AC$. Это последнее неравенство показывает, что наименьшее возможное значение суммы $a + b + c$ получилось бы, если бы P совпадало с A, что противоречит сделанному допущению, что P отлично от A, B, C. Таким образом, доказано, что точки A и B находятся вне круга K. Точно таким же образом доказывается, что точки B, C находятся вне круга с центром A и радиусом a, а точки A, C — вне круга с центром B и радиусом b.

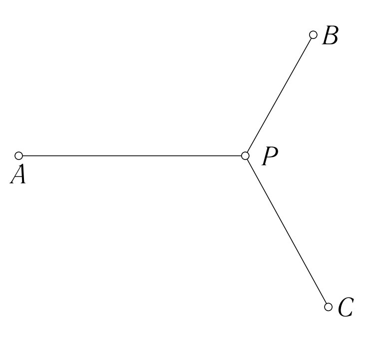


Рис. 208. Проблема Штейнера: $PA + PB + PC = \text{minimum}$

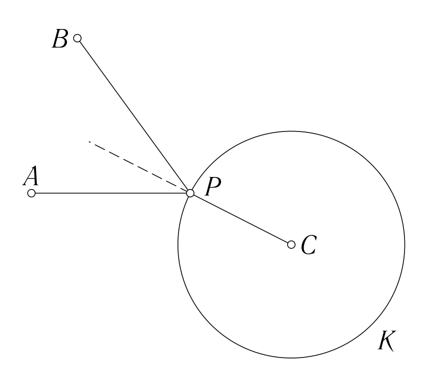


Рис. 209. К проблеме Штейнера

2. Анализ возникающих возможностей. Чтобы установить, которая именно из двух возможностей имеет место, нам придется обратиться к построению точки P. Для нахождения точки P, из которой две стороны треугольника, например AC и BC, видны под углом в $120°$, достаточно через точки A, C провести такую окружность K_1, у которой меньшая из дуг AC содержала бы $120°$, и через точки B, C провести окружность K_2, обладающую таким же свойством; затем взять точку пересечения двух дуг, содержащих по $120°$, если только эти дуги действительно пересекаются. Из точки P, найденной таким образом, сторона AB непременно также будет видна под углом в $120°$, так как сумма трех углов с вершиной P равна $360°$.

Из рис. 210 видно, что если все три угла треугольника ABC меньше $120°$, то две упомянутые дуги непременно пересекутся внутри треугольника. С другой стороны, если один из углов треугольника ABC, например C, больше чем $120°$, то дуги, о которых идет речь, не пересекутся (рис. 211). В этом случае не существует точки P, из которой каждая из трех сторон ABC была бы видна под углом в $120°$: окружности K_1 и K_2 пересекаются в точке P', из которой стороны AC и BC видны под углом в $60°$, и только одна сторона AB, противолежащая тупому углу, видна под углом в $120°$.

Если один из углов треугольника больше $120°$, то, как мы только что видели, нет такой точки P, из которой каждая из сторон видна под углом в $120°$; значит, искомая точка (в которой достигается минимум) должна совпадать с одной из вершин (так как это на основании § 1 — единственная иная возможность), а именно, с вершиной тупого угла.

Если же у треугольника все углы меньше $120°$, тогда, как мы видели, точку P, из которой все стороны видны под углом в $120°$, можно построить. Но, чтобы доказательство было закончено, нужно еще доказать, что для такой точки P сумма $a+b+c$ меньше, чем для любой из вершин треугольника (так как еще покамест неизвестно, которая из двух возможностей в рассматриваемом случае имеет место). Итак, докажем, например, что $a+b+c$ меньше, чем $AB+AC$ (рис. 212). С этой целью продолжим отрезок BP и спроектируем точку A на полученную прямую; пусть найденная проекция есть D. Так как, очевидно, $\angle APD = 60°$, то длина проекции PD равна $\frac{1}{2}a$. Так как BD есть проекция AB на прямую BP, то, значит, $BD < AB$. Но $BD = b + \frac{1}{2}a$; поэтому $b + \frac{1}{2}a < AB$. Совершенно таким же образом, проектируя A на продолжение отрезка PC, мы убеждаемся, что $c + \frac{1}{2}a < AC$. Складывая два последних неравенства, получаем: $a+b+c < AB+AC$. Итак, искомая точка не может находиться в вершине A. Так как, аналогично, она не может

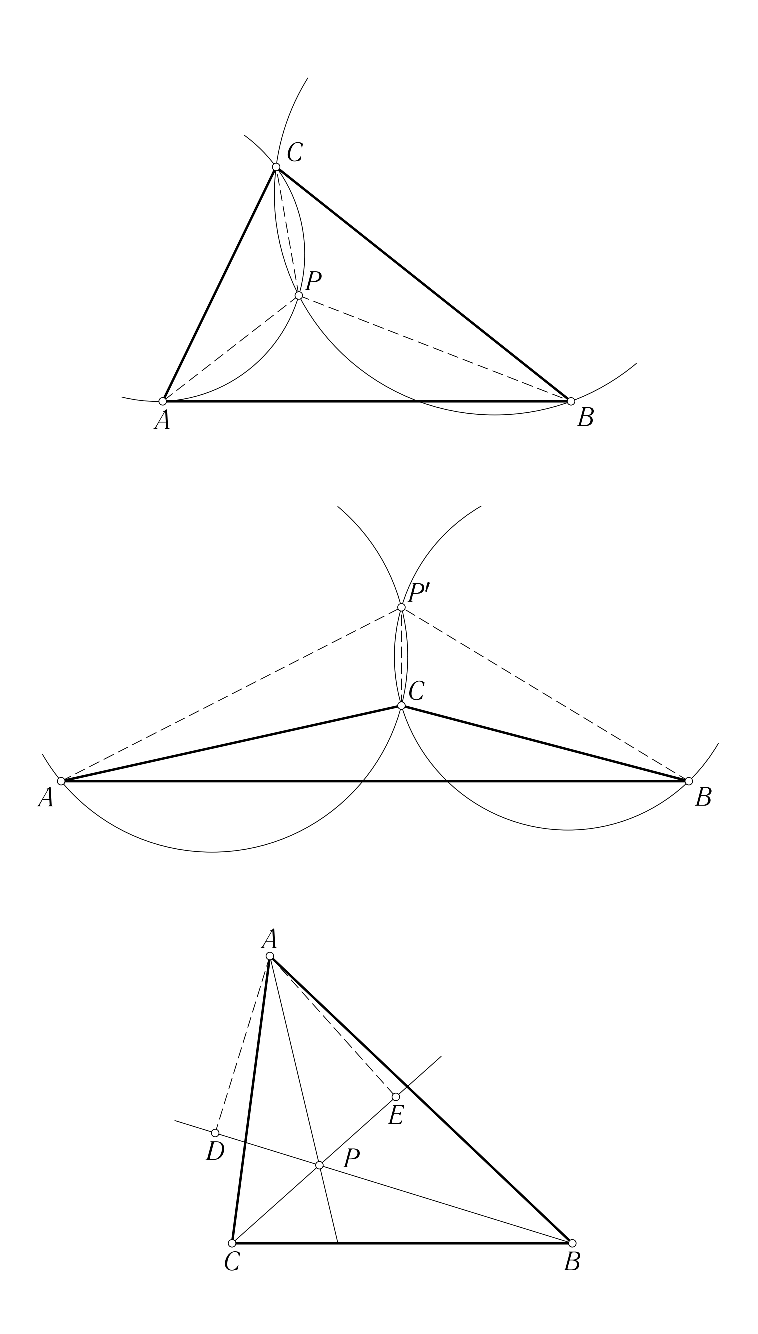


Рис. 210–212. К анализу различных возможностей в проблеме Штейнера

находиться также в вершинах B или C, то, следовательно, найденная точка P, из которой стороны видны под углом в $120°$, решает задачу.

3. Дополнительная проблема. Формальные математические методы нередко ведут дальше поставленных заранее целей. Так, если угол при вершине C больше $120°$, то вместо точки P (каковая совпадает на этот раз с точкой C) процедура геометрического построения дает другую точку P' — ту, из которой наибольшая сторона треугольника AB видна под углом в $120°$, а две другие стороны под углом в $60°$. Конечно, точка P' не дает решения рассматриваемой проблемы, но можно догадываться, что она имеет какое-то к ней отношение. Оказывается, в самом деле, что точка P' решает следующую проблему: минимизировать выражение $a + b - c$. Доказательство, вполне аналогичное изложенному выше для случая выражения $a + b + c$ и основанное на прямых результатах (§ 1, пункт 5), предоставляется в качестве упражнения читателю. Соединяя вместе полученные выводы, мы приходим к общей теореме.

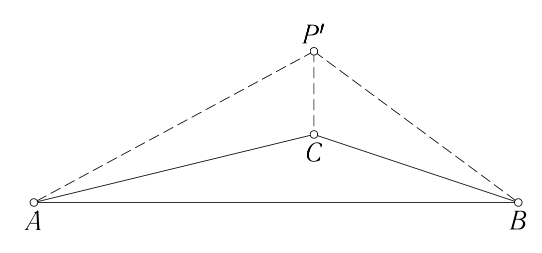


Рис. 213. Дополнительная проблема

Если все углы треугольника ABC меньше $120°$, то сумма $a + b + c$ расстояний a, b, c некоторой точки от точек A, B, C (соответственно) обращается в минимум в точке P, из которой каждая из сторон видна под углом в $120°$, а выражение $a + b - c$ обращается в минимум в вершине C; если же один из углов, скажем C, больше $120°$, то $a + b + c$ минимизируется в точке C, а $a + b - c$ — в точке P', из которой две меньшие стороны треугольника видны под углом в $60°$, а бóльшая — под углом в $120°$.

Таким образом, из двух минимальных проблем всегда одна решается построением окружностей, решение другой дается одной из вершин. В случае, когда $\angle C = 120°$, решения обеих проблем совпадают, так как точка, получаемая при геометрическом построении, оказывается вершиной C.

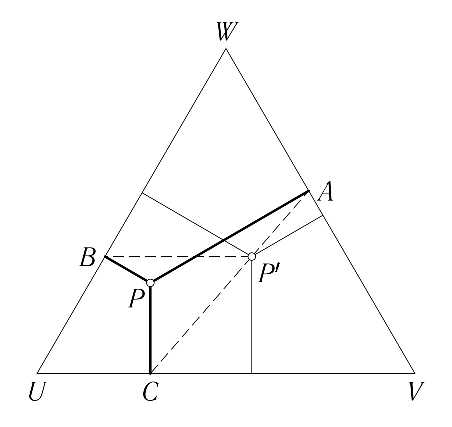


Рис. 214. Другое доказательство правильности решения Штейнера

4. Замечания и упражнения. Из произвольной точки P, взятой внутри равностороннего треугольника UVW, опустим перпендикуляры PA, PB, PC на три стороны (рис. 214). Тогда точки A, B, C и P образуют как раз такую фигуру, как мы рассматривали выше. Это замечание может быть использовано при решении проблемы Штейнера: достаточно, исходя из точек A, B, C, найти вершины равностороннего треугольника U, V, W.

Упражнения. 1) Выполните указанное построение, основываясь на том обстоятельстве, что сумма трех перпендикуляров, опущенных на стороны из произвольной точки P внутри равностороннего треугольника, постоянна, а именно, равна высоте треугольника.

2) Основываясь на аналогичном обстоятельстве в случае, когда P находится вне UVW, исследуйте дополнительную проблему.

В трехмерном пространстве можно рассмотреть проблему, подобную штейнеровской: по заданным четырем точкам A, B, C, D найти такую пятую точку P, чтобы сумма $a + b + c + d$ обращалась в минимум.

* **Упражнение.** Исследуйте эту трехмерную проблему и дополнительную к ней методами § 1 или же пользуясь правильным тетраэдром.

5. Обобщение: проблема уличной сети. В проблеме Штейнера были заданы три точки A, B, C. Было бы естественно обобщить эту проблему на случай n заданных точек A_1, A_2, ..., A_n следующим образом: требуется найти в плоскости такую точку P, чтобы сумма расстояний $a_1 + a_2 + \ldots + a_n$ (где a_i обозначает расстояние PA_i) обращалась в минимум. (В случае четырех точек, расположенных так, как показано на рис. 215, в качестве P нужно взять точку пересечения диагоналей четырехугольника $A_1A_2A_3A_4$; пусть читатель проверит это в качестве упражнения.) Эта обобщенная проблема, также изученная Штейнером, не ведет к интересным результатам. В данном случае мы имеем дело с поверхностным обобщением, подобных

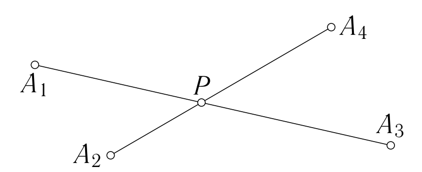


Рис. 215. Минимум суммы расстояний до четырех точек

которому немало встречается в математической литературе. Чтобы получить действительно достойное внимания обобщение проблемы Штейнера, приходится отказаться от поисков одной-единственной точки P. Вместо того поставим задачей построить «уличную сеть» или «сеть дорог между данными деревнями», обладающую минимальной общей длиной. Точнее, даны n точек A_1, A_2, ..., A_n; требуется найти такую связную систему прямолинейных отрезков, чтобы: 1) любые две из данных точек могли быть связаны ломаной линией, стороны которой входили бы в состав системы, 2) общая длина всей системы была наи-

меньшей.[1]

Решение этой задачи имеет тот или иной вид в зависимости от расположения данных точек. Читатель с пользой сможет заняться более внимательным рассмотрением этого вопроса, исходя из проблемы Штейнера. Мы ограничимся здесь указанием результатов в типических примерах, изображенных на рис. 216–218. В первом примере решение дается системой из пяти отрезков с двумя «кратными точками», в которых сходятся по три сегмента, образуя между собой углы в 120°. Во втором примере число кратных точек равно трем. При некоторых иных расположениях данных точек указанные фигуры не получаются: возможны случаи «вырождения», когда какая-нибудь одна из данных точек (или несколько таких точек) становится сама «кратной точкой» сети — таков третий из приведенных примеров.

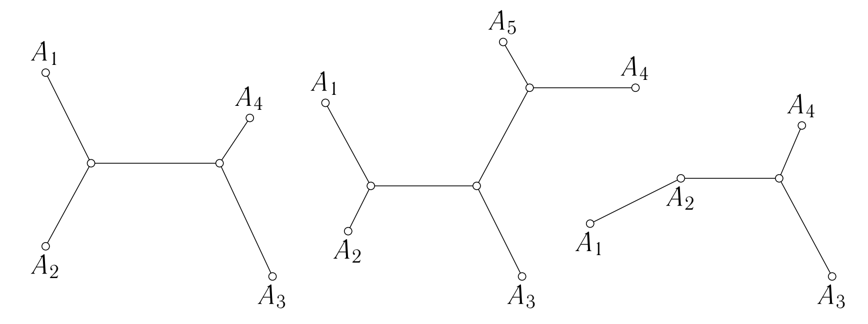


Рис. 216–218. Кратчайшая система путей, соединяющих данные точки

Если число данных точек равно n, то всего будет не более $n-2$ кратных точек, в которых сходятся по три отрезка, образуя углы в 120°.

Решение проблемы не всегда единственно. Так, если четыре данные точки расположены в вершинах квадрата, то возникают два эквивалентных решения (рис. 219, 220). Если точки A_1, A_2, ..., A_n являются вершинами ломаной линии с углами при вершинах, достаточно близкими к 180°, то сама ломаная является решением.

§ 6. Экстремумы и неравенства

Одной из характерных черт высших разделов математики является та выдающаяся роль, которую в них играют неравенства. В сущности, любая максимальная проблема всегда приводит к неравенству, выражающему тот факт, что рассматриваемая переменная величина не превышает некоторого максимального значения, доставляемого решением этой

---

[1] Выработка общих методов для решения прикладных задач типа описанной составила в последние годы предмет так называемого *линейного программирования*; см., например, [60]. — *Прим. ред.*

проблемы. Во многих случаях получаемые таким образом неравенства заслуживают внимания и независимо от экстремальной проблемы, к ним приводящей. В качестве примера мы рассмотрим сейчас важное неравенство, связывающее арифметическое и геометрическое средние.

1. Среднее арифметическое и среднее геометрическое двух положительных величин. Займемся прежде всего очень простой максимальной проблемой, с которой часто приходится встречаться и в самой математике, и в ее приложениях. В геометрической формулировке проблема эта заключается в следующем: среди всех прямоугольников с наперед заданным периметром найти тот, который имеет наибольшую площадь. Решением, как нетрудно догадаться, является квадрат. Доказать это можно следующим рассуждением. Пусть заданный периметр равен $2a$. Тогда сумма $x+y$ длин двух прилежащих сторон прямоугольника x и y равна постоянной величине a, а в максимум следует обратить произведение xy. «Среднее арифметическое» величин x и y есть не что иное, как выражение

$$m = \frac{x+y}{2}.$$

Введем еще величину

$$d = \frac{x-y}{2},$$

причем получатся соотношения

$$x = m + d, \qquad y = m - d;$$

из них вытекает, что

$$xy = (m+d)(m-d) = m^2 - d^2 = \left(\frac{x+y}{2}\right)^2 - d^2.$$

Так как d^2 не может быть отрицательно, а обращается в нуль только при $x = y$, то мы немедленно приходим к неравенству

$$\sqrt{xy} \leqslant \frac{x+y}{2}, \tag{1}$$

причем знак равенства здесь возможен только при $d = 0$, т. е. при $x = y$.

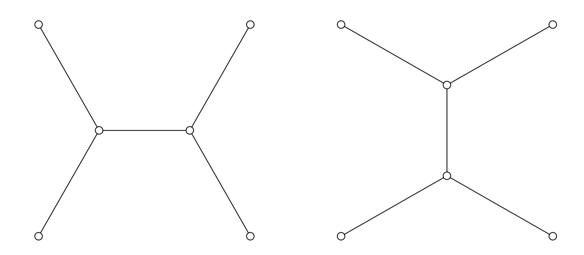


Рис. 219–220. Кратчайшие системы путей, соединяющие вершины квадрата

Так как $x + y$ имеет постоянное значение a, то отсюда следует, что выражение $\sqrt{xy}$, а значит, и интересующая нас площадь xy принимают наибольшие возможное значение при $x = y$. Выражение

$$g = \sqrt{xy},$$

где радикал взят в арифметическом смысле — со знаком +,— называется «средним геометрическим» положительных величин x и y; неравенство (1) выражает основное соотношение между средними арифметическим и геометрическим.

Неравенство (1) вытекает также непосредственно из того факта, что выражение

$$(\sqrt{x} - \sqrt{y})^2 = x + y - 2\sqrt{xy},$$

будучи точным квадратом, не может быть отрицательным и обращается в нуль только при $x = y$.

Вот еще геометрический вывод того же неравенства. Рассмотрим в плоскости x, y неподвижную прямую линию $x + y = 2m$ и вместе с ней семейство кривых (гипербол) $xy = c$, причем c постоянно для каждой кривой, но меняется при переходе от одной кривой к другой. Из рис. 221 ясно, что кривой, имеющей хоть одну общую точку с нашей прямой линией и соответствующей наибольшему значению c, является та гипербола, которая касается прямой в точке $x = y = m$; для этой гиперболы, следовательно, $c = m^2$. Итак,

$$xy \leqslant \left(\frac{x+y}{2}\right)^2.$$

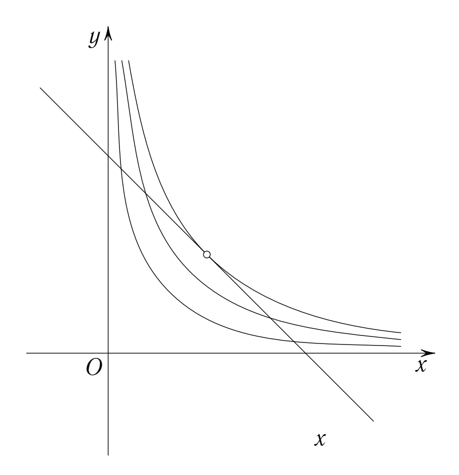


Рис. 221. Минимум xy при заданном значении $x + y$

Следует заметить, что всякое неравенство вида $f(x, y) \leqslant g(x, y)$ можно прочесть двумя способами, и потому оно порождает как максимальное, так и минимальное свойства. Например, неравенство (1) выражает также тот факт, что среди всех прямоугольников с данной площадью именно квадрат имеет наименьший периметр.

2. Обобщение на случай n переменных. Неравенство (1), связывающее средние арифметическое и геометрическое двух положительных величин, может быть обобщено на любое число n положительных величин, которые мы будем обозначать $x_1, x_2, x_3, \ldots, x_n$. Средним арифметическим этих величин называют выражение

$$m = \frac{x_1 + x_2 + \ldots + x_n}{n},$$

а средним геометрическим — выражение

$$g = \sqrt[n]{x_1 x_2 \ldots x_n},$$

причем имеется в виду всегда положительное значение радикала. Общая теорема утверждает, что

$$g \leqslant m \tag{2}$$

и что равенство $g = m$ возможно только в том случае, если все величины x_i равны между собой.

Было предложено много различных остроумных доказательств этого общего результата. Простейший метод заключается в применении того же простого рассуждения, которое мы провели в пункте 1. Перед нами стоит проблема: разбить данное положительное число C на n положительных слагаемых, $C = x_1 + x_2 + \ldots + x_n$, таким образом, чтобы произведение $P = x_1 x_2 \ldots x_n$ было возможно большим. Мы будем исходить из допущения, на первый взгляд очевидного, но мы позднее будем иметь случай его проанализировать (§ 7), что наибольшее значение P существует и достигается, скажем, при значениях $x_1 = a_1$, $x_2 = a_2$, ..., $x_n = a_n$. Нам достаточно установить, что $a_1 = a_2 = \ldots = a_n$, ибо в этом случае $g = m$. Допустим, что это не так: пусть, например, $a_1 \neq a_2$. Тогда рассмотрим значения

$$x_1 = s, \quad x_2 = s, \quad x_3 = a_3, \ \ldots, \ x_n = a_n,$$

где

$$s = \frac{a_1 + a_2}{2}.$$

Другими словами, мы заменим прежнюю систему значений величин x_i новой системой, которая отличается от прежней лишь тем, что значения двух первых величин x_1 и x_2 сделаны равными между собой, причем общая сумма C остается неизменной. Мы можем написать

$$a_1 = s + d, \qquad a_2 = s - d,$$

где положено

$$d = \frac{a_1 - a_2}{2}.$$

Новое произведение равно

$$P' = s^2 \cdot a_3 \ldots a_n,$$

тогда как прежнее произведение было

$$P = (s + d)(s - d) \cdot a_3 \ldots a_n = (s^2 - d^2) \cdot a_3 \ldots a_n.$$

Отсюда ясно, что при $d \neq 0$

$$P < P',$$

а это противоречит сделанному допущению, что произведение P имеет максимальное значение. Итак, $d = 0$, и тогда $a_1 = a_2$. Таким же образом

доказывается, что $a_1 = a_i$, где a_i обозначает любое из чисел a; отсюда следует, что все числа a равны между собой. Мы убедились в том, что 1) $g = m$, если все числа x_i равны между собой, 2) наибольшее значение g получается только тогда, когда все числа x_i равны между собой. Отсюда можно заключить, что во всех прочих случаях $g < m$. Теорема доказана.

3. Метод наименьших квадратов. Среднее арифметическое n чисел $x_1, x_2, \ldots, x_n$ (которые здесь нет необходимости считать обязательно положительными) обладает замечательным минимальным свойством. Пусть u — числовое значение некоторой неизвестной величины, которое мы хотим определить насколько возможно точнее с помощью какого-то измерительного инструмента. Пусть произведено для этой цели n измерений, которые дали результаты $x_1, x_2, \ldots, x_n$, слегка различающиеся между собой, что обусловливается неизбежными и зависящими от разных причин измерительными ошибками. Возникает вопрос: какое же значение следует приписать величине u в качестве заслуживающего наибольшего доверия? В качестве «истинного» или «оптимального» значения принято выбирать среднее арифметическое

$$m = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Дать подлинное обоснование указанной процедуре было бы невозможно, не углубляясь в пространные рассуждения, относящиеся к области теории вероятностей. Все же мы можем здесь отметить некоторое минимальное свойство средней арифметической m, которое до некоторой степени оправдывает ее выбор. Пусть u — какое угодно числовое значение измеряемой величины. Тогда разности $u - x_1, u - x_2, \ldots, u - x_n$ представляют собой отклонения этой величины от результатов отдельных наблюдений. Эти отклонения могут быть частью положительными, частью отрицательными, и совершенно естественно стремиться к такому оптимальному выбору u, при котором «тотальное» (в каком-то смысле) отклонение было бы возможно меньше. Следуя Гауссу, берут обыкновенно в качестве «измерителей неточности» не сами отклонения, а их квадраты $(u - x_i)^2$ и затем выбирают оптимальное значение u с таким расчетом, чтобы минимизировать «тотальное» отклонение, под каковым понимают сумму квадратов отдельных отклонений

$$(u - x_1)^2 + (u - x_2)^2 + \ldots + (u - x_n)^2.$$

Определенное таким образом оптимальное значение u есть не что иное, как среднее арифметическое m: в этом заключается исходное положение знаменитого «метода наименьших квадратов», созданного Гауссом. Мы постараемся возможно проще доказать это утверждение. Если мы напишем

$$(u - x_i) = (m - x_i) + (u - m),$$

то получим

$$(u-x_i)^2=(m-x_i)^2+(u-m)^2+2(m-x_i)(u-m).$$

Сложим, далее, все такие равенства, полагая $i=1, 2, \ldots, n$. Последний член при этом дает

$$2(u-m)(nm-x_1-\ldots-x_n),$$

а это выражение по определению m равно нулю. Следовательно, мы получаем:

$$(u-x_1)^2+(u-x_2)^2+\ldots+(u-x_n)^2= \\ =(m-x_1)^2+(m-x_2)^2+\ldots+(m-x_n)^2+n(u-m)^2.$$

Отсюда уже ясно, что

$$(u-x_1)^2+(u-x_2)^2+\ldots+(u-x_n)^2\geqslant \\ \geqslant(m-x_1)^2+(m-x_2)^2+\ldots+(m-x_n)^2,$$

причем знак равенства возможен только при $u=m$. Как раз это самое мы и собирались доказать.

Общий метод наименьших квадратов принимает руководящий принцип — минимизировать сумму квадратов — во всех более сложных случаях, когда нужно как-то согласовать между собой ряд слегка противоречащих друг другу данных наблюдения. Так, представим себе, что измерены координаты x_i, y_i для n точек, которые теоретически должны лежать на прямой линии, и предположим, что полученные таким эмпирическим путем точки оказываются расположенными по прямой не вполне точно. Как выбрать прямую, которая наилучшим образом была бы «приложена» или «подогнана» к этим точкам? Руководящий принцип приводит к следующему приему (который — необходимо признать — мог бы быть заменен и другими процедурами, основанными на иных рассуждениях). Пусть $y=ax+b$ есть уравнение искомой прямой, так что наша проблема заключается в определении коэффициентов a и b. Измеренное по направлению оси y расстояние прямой от точки x_i, y_i равно $y_i-(ax_i+b)$, т. е. y_i-ax_i-b, причем имеет положительный или отрицательный знак, смотря по тому, расположена ли точка выше или ниже прямой. Тогда квадрат этого расстояния равен $(y_i-ax_i-b)^2$, и согласно основному принципу метода наименьших квадратов нам достаточно подобрать a и b таким образом, чтобы выражение

$$(y_1-ax_1-b)^2+(y_2-ax_2-b)^2+\ldots+(y_n-ax_n-b)^2$$

достигало наименьшего возможного значения. Мы приходим, таким образом, к минимальной проблеме с двумя переменными величинами a и b. Хотя решение этой проблемы с исследованием всех подробностей и не представляет особенной трудности, мы все же воздержимся здесь от его рассмотрения.

§ 7. Существование экстремума. Принцип Дирихле

1. Общие замечания. В некоторых из рассмотренных нами экстремальных проблем прямо доказывалось, что решение дает наилучший результат из числа прочих возможных. Ярким примером является принадлежащее Шварцу решение задачи о треугольнике: здесь сразу видно, что никакой вписанный треугольник не может иметь меньший периметр, чем высотный треугольник. Некоторые примеры такого же типа связаны с явно написанными неравенствами, каково, например, неравенство между средними арифметическим и геометрическим. Но при решении других проблем мы шли по иному пути. Мы допускали прежде всего, что решение уже найдено, и затем, анализируя это допущение, получали заключения, иногда позволяющие дать полную характеристику решения и выполнить соответствующее его построение. Так именно обстояло дело с проблемой Штейнера и таков же был план второго решения проблемы Шварца. Названные два метода логически различны. Первый метод, пожалуй, можно считать более совершенным, так как он дает конструктивное доказательство правильности результата. Второй метод, если судить по примеру проблемы Шварца (второе решение), кажется более простым. Но он является не прямым, а косвенным и, самое главное, он условен по самой своей структуре, так как предполагает *существование* решения. Он приводит к окончательному результату лишь постольку, поскольку существование решения или постулировано, или доказано. Без этой предпосылки он показывает всего-навсего, что *если* решение существует, то оно обладает такими-то свойствами.[1]

Вследствие кажущейся очевидности предпосылки о существовании решения математики вплоть до конца прошлого столетия не обращали особенного внимания на указанное логическое обстоятельство и допускали существование решения экстремальных проблем как нечто само собой разумеющееся. Некоторые из величайших ученых XIX в. — Гаусс, Дирихле, Риман — некритически основывали на такого рода допущении глубокие и иначе трудно доказуемые предложения в области математической физики и теории функций. Кризис наступил вскоре после того, как Риман в 1849 г. опубликовал свою докторскую диссертацию, посвященную основаниям теории функций комплексного переменного. Эта сжато написанная работа, представляющая собой один из величайших подвигов математической мысли в новейшую эпоху, была до такой

[1] Логическая необходимость доказывать существование экстремума иллюстрируется следующим парадоксом: 1 есть наибольшее целое число. Вот доказательство. Пусть x есть наибольшее целое число. Если допустим, что $x > 1$, то отсюда следует $x^2 > x$, что противоречит сделанному допущению. Итак, x должен быть равен 1.

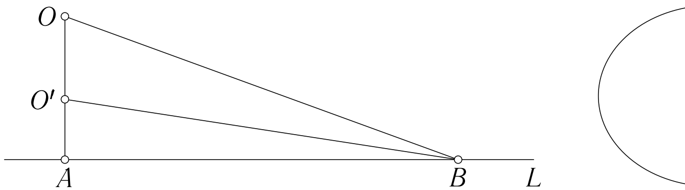
степени неортодоксальной в трактовке вопроса, что многие предпочли попросту ее игнорировать. Вейерштрасс в то время был уже знаменитым профессором Берлинского университета и пользовался репутацией основателя строго построенной теории функций. Вначале пораженный, но все же исполненный сомнений, он вскоре обнаружил в работе Римана логическую брешь, о восполнении которой сам автор не позаботился. Уничтожающая критика Вейерштрасса не поколебала уверенности Римана в справедливости полученных им результатов, но долгое время его теория не пользовалась признанием. Головокружительная научная карьера Римана вскоре внезапно оборвалась: он умер от туберкулеза. Все же его идеи поддерживались и дальше несколькими убежденными и преданными учениками. Только через пятьдесят лет после появления диссертации Римана Гильберту удалось наконец открыть пути, приводящие к исчерпывающему ответу на все вопросы, оставленные в стороне и не разрешенные Риманом. Начатое Риманом и развернувшееся во второй половине столетия развитие математических теорий, глубоко проникающих в область физики, представляет одну из самых блестящих страниц в истории современной науки.

Уязвимое место в работе Римана — как раз вопрос о существовании минимума. Свою теорию Риман основывает на принципе Дирихле (так он сам его назвал по имени своего учителя: Дирихле, читая лекции в Гёттингене, пользовался этим принципом, но ни в одной из своих работ о нем не писал). Предположим, для большей определенности, что некоторая часть плоскости или какой-нибудь поверхности покрыта слоем станиоля и что стационарный электрический ток проходит по слою, соединенному в двух точках с полюсами батареи. Нет сомнений, что такой эксперимент приведет к некоторому однозначно определенному распределению токов. Но как обстоит дело с соответствующей математической проблемой — проблемой, имеющей первостепенное значение в теории функций и в других областях? В теории электричества рассматриваемое нами физическое явление описывается как «дифференциальное уравнение в частных производных с граничными условиями». Именно эта математическая проблема нас и интересует; возможность ее решения кажется правдоподобной именно по той причине, что мы допустили ее эквивалентность физическому явлению; но математическое доказательство этой возможности никоим образом не может базироваться на сделанном допущении. В подходе Римана к решению рассматриваемого им математического вопроса можно различить два этапа. Во-первых, он показывает, что проблема эквивалентна некоторой минимальной проблеме: некоторая величина, выражающая энергию потока электричества, минимизируется некоторыми реально осуществляющимся потоком (по сравнению с иными потоками, совместными с предписанными граничными условиями). Во-вторых, в качестве «принципа Дирихле» он вводит

постулат о том, что такого рода минимальная проблема допускает решение. Риман решительно ничего не предпринял для того, чтобы найти хоть какое-нибудь математическое оправдание для этого постулата, и именно в этом пункте его настигли атаки со стороны Вейерштрасса. Не только существование минимума само по себе не было очевидным, но, как выяснилось впоследствии, вопрос оказался чрезвычайно тонким: математика того времени еще не была подготовлена к его решению, и только через несколько десятилетий напряженные усилия исследовательской мысли привели к законченным результатам.

2. Примеры. Мы проиллюстрируем возникающую трудность двумя примерами.

1) Отметим на прямой L две точки A и B, находящиеся на расстоянии d, и поставим задачу — отыскать ломаную линию кратчайшей длины, которая, выходя из точки A по направлению, перпендикулярному к L, заканчивалась бы в точке B. Так как прямолинейный отрезок AB безусловно короче всех других путей, связывающих точки A и B, то можно быть уверенным, что любой допустимый (удовлетворяющий требованиям задачи) путь имеет длину большую чем d: в самом деле, единственный путь, длина которого равна d, есть прямолинейный отрезок AB, а он не удовлетворяет требованию относительно направления в точке A, т. е. не является допустимым. Рассмотрим, с другой стороны, допустимый путь AOB на рис. 222. Заменяя точку O точкой O', расположенной ближе к A, мы можем получить новый допустимый путь, длина которого как угодно мало отличается от d; значит, если существует *кратчайший* допустимый путь, то длина его не может быть больше, чем d и, следовательно, должна быть в точности равна d. Но мы видели, что единственный путь, имеющий такую длину, не является допустимым. Итак, кратчайшего допустимого пути не существует, и задача наша не имеет решения.



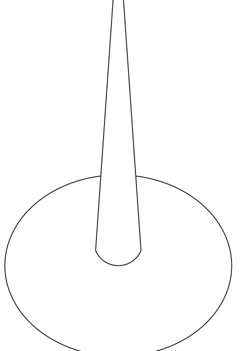


Рис. 222–223. К вопросу о существовании минимума

2) Пусть C — некоторый круг, а S — точка, лежащая выше его центра на расстоянии 1 (рис. 223). Рассмотрим множество поверхностей, ограни-

ченных окружностью C и проходящих через точку S, притом лежащих «над» кругом C (в том смысле, что никакие две различные точки этой поверхности не могут вертикально проектироваться в одну и ту же точку круга C). Какая поверхность из рассматриваемого множества обладает наименьшей площадью? Каким бы естественным ни казался этот вопрос, положительного ответа на него дать нельзя: допустимой поверхности с наименьшей площадью не существует. Если бы поверхность не была подчинена условию проходить через точку S, тогда решением задачи был бы, очевидно, плоский диск, ограниченный окружностью C. Обозначим площадь этого диска через A. Всякая другая поверхность, ограниченная окружностью C, непременно имеет площадь бо́льшую, чем A. Но можно указать допустимую поверхность, площадь которой будет отличаться от A как угодно мало. В самом деле, возьмем коническую поверхность высоты 1 — такую «тоненькую», чтобы ее площадь была меньше заранее назначенного маленького числа. Поместим эту поверхность посреди диска так, чтобы ее вершина попала в точку S, и затем рассмотрим поверхность, образованную из нашей конической поверхности и той части диска, которая окажется вне основания конуса. Совершенно ясно, что построенная таким образом поверхность, только вблизи центра диска отличающаяся от самого диска, обладает площадью, превосходящей A меньше чем на заранее назначенное число. Так как это последнее число может быть выбрано сколь угодно малым, то отсюда следует, что минимум площади (если он существует) не может отличаться от площади диска A. Но среди поверхностей, ограниченных контуром C, только сам диск обладает площадью A; однако диск не проходит через точку S и, значит, не является допустимой поверхностью; следовательно, решения задачи не существует.

Мы можем избавить себя от труда приводить дальнейшие относящиеся сюда софистически утонченные примеры, указанные Вейерштрассом. Уже приведенные примеры достаточно убедительно показывают, что существование минимума не является тривиальным моментом в математическом доказательстве. Попробуем взглянуть на рассматриваемый вопрос с более отвлеченной точки зрения. Представим себе некоторый определенный класс объектов, например, кривых или поверхностей; пусть каждому объекту этого класса поставлено в соответствие — как функция этого объекта — некоторое число, например длина или площадь. Если в классе содержится лишь конечное число объектов, то среди соответствующих чисел неизбежно имеется наибольшее и наименьшее. Но если в классе содержится бесконечное множество объектов, то даже в том случае, если все соответствующие числа заключены между двумя конечными границами, среди них вовсе не обязательно найдется наибольшее и наименьшее. На числовой оси множество чисел изображается в виде множества точек. Предположим, ограничившись простейшим

случаем, что все числа множества положительные. Такое множество непременно имеет «нижнюю границу» — такое число α, меньше которого в нашем множестве нет ни одного числа и которое или само есть элемент множества, или как угодно мало отличается от некоторого элемента множества. Если α само принадлежит множеству, то оно является его наименьшим элементом; в противном случае множество не содержит вовсе наименьшего элемента. Например, множество чисел $1, \frac{1}{2}, \frac{1}{3}, \ldots$ не содержит наименьшего элемента, так как нижняя граница 0 не принадлежит множеству. Такого рода отвлеченные примеры иллюстрируют логические трудности, связанные с проблемой существования. Математическое решение минимальной проблемы нельзя назвать исчерпывающим, если в явной или в неявной форме не устанавливается, что среди элементов числового множества, рассматриваемого в связи с проблемой, существует наименьший.

3. Экстремальные проблемы элементарного содержания. В задачах элементарного содержания бывает достаточно внимательно проанализировать условия, чтобы уяснить, как обстоит дело с существованием решения. В главе VI, § 5, было исследовано общее понятие компактного множества и было установлено, что непрерывная функция, заданная на некотором множестве элементов, для каких-то элементов множества непременно достигает своих экстремальных значений, если данное множество обладает свойством компактности. В любой из вышеприведенных элементарных проблем сравниваемые между собой числовые элементы могли быть рассматриваемы как значения функции одной или нескольких переменных в области, которая или была компактным множеством, или — без существенного видоизменения проблемы — могла быть сделана таковым. В таких случаях существование максимума или минимума не подлежало сомнению. Остановимся, в качестве примера, на проблеме Штейнера. Рассматриваемая в ней величина есть сумма трех расстояний, и эта последняя зависит от положения точки непрерывно. Хотя область, в которой может двигаться точка, есть вся плоскость, мы можем без ограничения общности провести окружность большого радиуса (включающую весь рисунок) и подчинить точку условию находиться внутри этой окружности или на ней самой. В самом деле, если движущаяся точка будет находиться достаточно далеко от вершин треугольника, сумма трех расстояний от сторон наверняка превысит $AB + AC$, а последняя величина принадлежит к числу подлежащих сравнению значений нашей функции. Таким образом, если существует минимум для «ограниченной» проблемы (когда точка подчинена дополнительному ограничению), то существует минимум и для неограниченной проблемы. С другой стороны, нетрудно удостовериться, что множество, состоящее из точек внутри круга или

на его границе, компактно. Итак, существование минимума в случае проблемы Штейнера доказано.

Насколько существенно свойство компактности области, в которой изменяется независимое переменное, обнаруживает следующий пример. Если заданы две замкнутые кривые C_1 и C_2, то всегда можно найти на C_1 и C_2 соответственно две такие точки P_1 и P_2, что расстояние между ними минимально, и можно найти две такие точки Q_1 и Q_2, что расстояние между ними максимально. Действительно, расстояние между точкой A_1 на C_1 и точкой A_2 на C_2 есть непрерывная функция, заданная на компактном множестве, элементы которого — пары точек A_1, A_2. Напротив, если данные кривые, не будучи замкнутыми, уходят в бесконечность, проблема может и не иметь решения. На рис. 224 изображены две такие кривые, что ни наименьшее, ни наибольшее расстояния между соответственно принадлежащими им точками не достигаются: при этом нижняя граница расстояний равна нулю, а верхняя граница бесконечна. В иных случаях существует минимум, но не существует максимума. Так, в случае двух ветвей гиперболы (рис. 17, стр. 96) минимальное расстояние реализуется для вершин A и A', тогда как нельзя указать пары точек, между которыми расстояние было бы максимальным.

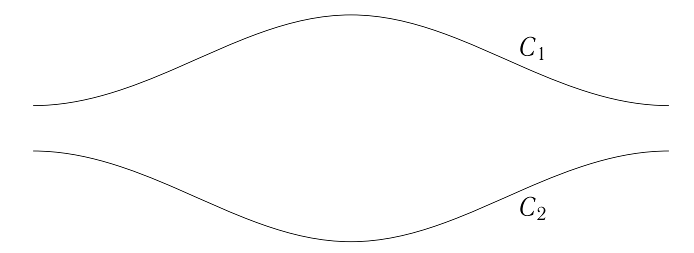


Рис. 224. Кривые, между которыми нет ни наименьшего, ни наибольшего расстояния

Нетрудно понять, чем обусловливается различие между двумя предыдущими примерами; для этого достаточно искусственно ограничить область изменения переменных. Возьмем произвольное положительное число R и подчиним абсциссы точек ограничению $|x| \leqslant R$. Тогда для обеих проблем будет существовать и минимум и максимум. Но в первом примере и минимум и максимум достигаются на границе области, каково бы ни было R, и при неограниченном возрастании R соответствующие точки удаляются в бесконечность. Напротив, во втором примере минимальное расстояние достигается внутри области, и точки, его реализующие, остаются неподвижными, как бы ни возрастало R.

4. Трудности, возникающие в более сложных случаях. Если вопрос о существовании экстремума не представляет серьезных за-

труднений в элементарных проблемах, зависящих от одной, двух или, вообще, конечного числа переменных, то дело обстоит совсем иначе в случае проблемы Дирихле или даже в случае более простых проблем такого же типа. Причина кроется или в том, что область изменения независимого переменного оказывается некомпактной, или в том, что рассматриваемая функция не является непрерывной. В первом примере пункта 2 мы имеем множество путей $AO'B$, причем O' стремится к A. Все такие пути, с точки зрения условия проблемы, одинаково допустимы. Но пути $AO'B$ в пределе переходят в прямолинейный отрезок AB, который сам уже не представляет собой допустимого пути. Множество допустимых путей в этом примере подобно множеству чисел $0 < x \leqslant 1$, для которого не имеет места теорема Вейерштрасса об экстремальных значениях (см. стр. 336). Точно такое же положение вещей и во втором примере: если конусы становятся все тоньше и тоньше, последовательность соответствующих поверхностей в пределе переходит в диск с перпендикуляром, торчащим вверх и заканчивающимся точкой S. Но этот предельный геометрический образ уже не может быть причислен к «допустимым» поверхностям: множество «допустимых» поверхностей и на этот раз не оказывается компактным.

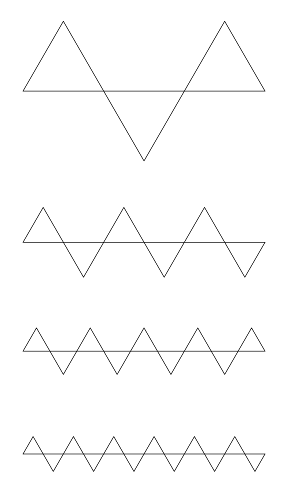


Рис. 225. Приближение отрезка ломаными линиями

В качестве примера зависимости, не обладающей свойством непрерывности, рассмотрим длину кривой. Длину кривой нельзя считать функцией от конечного числа числовых переменных, так как кривая в целом не может быть охарактеризована конечным числом «координат», и зависимость длины кривой от самой кривой не является непрерывной. Чтобы убедиться в этом, соединим две точки A и B, отстоящие одна от другой на расстоянии d, зигзагообразной ломаной P_n, вместе с отрезком AB образующей n равносторонних треугольников. Из рис. 225 ясно видно, что длина P_n при любом n равна в точности $2d$. Рассмотрим теперь последовательность ломаных линий P_1, P_2, ... Отдельные зигзаги ломаной линии P_n уменьшаются по своей высоте, в то время как число их увеличивается, и совершенно ясно, что ломаная P_n в пределе переходит в прямолинейный отрезок AB, в котором уже нет и следов «зигзагообразности». Но длина P_n все время равна $2d$, каково бы ни было n, тогда как длина предельного отрезка AB равна всего лишь d. Длина кривой, таким образом, не зависит «от самой кривой» непрерывно.

Все приведенные примеры подтверждают, что при исследовании во-

проса о существовании решения экстремальных проблем в более сложных случаях следует проявлять крайнюю осмотрительность.

§ 8. Изопериметрическая проблема

Что среди всех замкнутых кривых данной длины именно окружность охватывает наибольшую площадь, — это один из «очевидных» фактов математики, строгое доказательство которых возможно только на основе новейших методов. Несколько остроумных способов доказательства этой теоремы предложил Штейнер; мы рассмотрим одно из его доказательств.

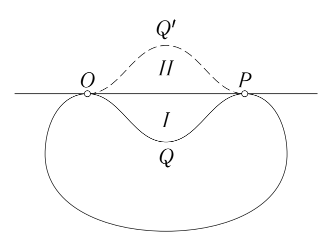


Рис. 226. К доказательству решения изопериметрической проблемы

Начнем с допущения, что решение проблемы существует. Приняв это, предположим, что это решение осуществляется некоторой кривой C, имеющей длину L и охватывающей максимальную площадь. Легко доказать, что кривая C выпуклая: это значит, что прямолинейный отрезок, соединяющий любые две точки C, лежит целиком внутри или на C. Если бы кривая C не была выпуклой, то, как показано на рис. 226, можно было бы указать отрезок OP, конечные точки которого находились бы на C, а сам он был бы вне C. Дуга $OQ'P$ — отражение дуги OQP относительно OP — образовывала бы вместе с дугой ORP кривую длины L, охватывающую площадь бо́льшую, чем охватывает данная кривая C, так как включала бы дополнительно площади I и II. Это противоречило бы допущению, что при данной длине L кривая C охватывает наибольшую площадь. Итак, кривая C должна быть выпуклой. Возьмем теперь какие-нибудь две точки A, B, которые делят кривую C (являющуюся решением проблемы) на две дуги равной длины. Тогда отрезок AB разделит область, ограниченную кривой C, на две равновеликие области.

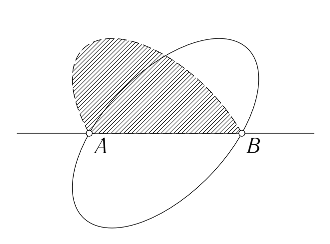


Рис. 227. К доказательству решения изопериметрической проблемы

В самом деле, если бы площади двух областей не были равны, то область большей площади можно было бы отразить относительно AB (рис. 227), и тогда получилась бы замкнутая кривая длины L, охватывающая площадь большую, чем та, которую охватывает кривая C. Отсюда следует, что любая незамкнутая кривая, представляющая собой половину (по длине) кривой C, является решением следующей пробле-

мы: найти дугу длины $\frac{L}{2}$ с конечными точками A, B, охватывающую вместе с отрезком AB максимальную площадь. Мы покажем теперь, что решением этой новой проблемы является полуокружность, и тогда будет ясно, что решением основной проблемы является окружность. Итак, пусть дуга AOB решает новую проблему. Достаточно убедиться в том, что всякий вписанный угол, например $\angle AOB$ (рис. 228), будет прямым: отсюда будет вытекать, что дуга AOB — полуокружность. Допустим, напротив, что угол AOB не прямой. Заменим тогда треугольник AOB другим треугольником с теми же сторонами AO и OB, но с заключенным между ними углом в 90°; тогда длина дуги AOB останется та же $\left(\frac{L}{2}\right)$, и притом заштрихованные фигуры не изменятся. Но площадь треугольника AOB при этом увеличится, так как треугольник с двумя данными сторонами имеет максимальную площадь при условии, что заключенный между ними угол — прямой (см. стр. 352). Итак, новая дуга AOB (рис. 229) вместе с отрезком AB охватит бóльшую площадь, чем первоначальная. Полученное противоречие приводит к заключению, что, какова бы ни была точка O на рассматриваемой дуге AB, угол AOB должен быть прямым. В таком случае доказательство можно считать законченным: кривая, решающая изопериметрическую проблему, есть окружность.

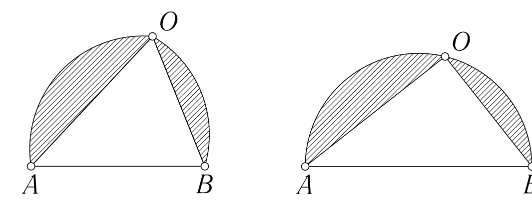


Рис. 228–229. К доказательству решения изопериметрической проблемы

Изопериметрическое свойство окружности может быть выражено в форме неравенства. Если L есть длина окружности, то охватываемая ею площадь равна $\frac{L^2}{4\pi}$, и потому, какова бы ни была замкнутая кривая, непременно оправдывается следующее *изопериметрическое неравенство*, связывающее длину кривой C и охватываемую ею площадь A:

$$A \leqslant \frac{L^2}{4\pi}.$$

Равенство здесь имеет место только в случае окружности.

* Как ясно из соображений, приведенных в § 7, доказательство Штейнера имеет лишь условное значение: «Если существует кривая C длины L, охватывающая максимальную площадь, то эта кривая —

окружность». Чтобы установить справедливость указанной предпосылки, нужна существенно иная аргументация. Прежде всего установим теорему элементарного содержания: среди всевозможных замкнутых многоугольников P_n с четным числом сторон $2n$ и обладающих периметром заданной длины наибольшую площадь имеет правильный $2n$-угольник. Доказательство строится по тому же образцу, что и приведенное выше доказательство Штейнера, со следующими изменениями. С вопросом о существовании решения здесь трудностей не возникает: $2n$-угольник, а также его периметр и площадь, зависит непрерывно от $4n$ координат его вершин, и, не ограничивая общности, область изменения этих координат (в $4n$-мерном пространстве) можно сделать компактной. Таким образом, мы можем смело начинать с утверждения, что некоторый $2n$-угольник P есть решение рассматриваемой теперь проблемы, и затем переходить к анализу его свойств. Как и в штейнеровском доказательстве, доказывается, что многоугольник P выпуклый. Затем убедимся, что все $2n$ сторон P равны между собой. Допустим, напротив, что две смежные стороны AB и BC имеют различные длины; тогда можно от многоугольника P отрезать треугольник ABC и заменить его равнобедренным треугольником $AB'C$, в котором $AB' + B'C = AB + BC$ и площадь которого больше (см. § 1). Тогда мы получим многоугольник P' с тем же периметром, но с большей площадью, вопреки сделанному допущению. Итак, все стороны P должны быть равны между собой. Остается показать, что многоугольник P правильный: для этого достаточно убедиться, что около P можно описать окружность. Доказательство строится дальше, как у Штейнера. Устанавливаем прежде всего, что всякая диагональ, соединяющая противоположные вершины, делит площадь на две равные части. Затем доказываем, что все вершины одного из многоугольников, возникающего при разрезании по диагонали, лежат на одной и той же окружности. Восстановить подробности намеченных доказательств (следующих образцу Штейнера) предоставляем читателю в качестве упражнения.

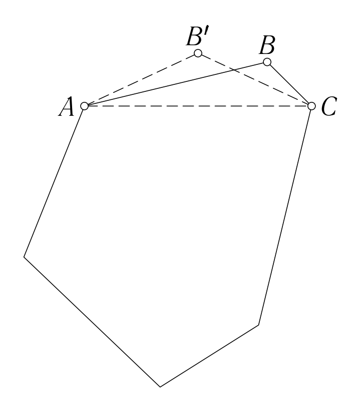


Рис. 230. К доказательству решения изопериметрической проблемы

Существование решения изопериметрической проблемы доказывается с помощью предельного перехода: когда мы увеличиваем неограниченно число сторон $2n$ многоугольника P, он в пределе переходит в окружность. Этот же предельный переход дает, очевидно, и само решение.

Рассуждение Штейнера непригодно для доказательства изопериметрического свойства сферы в трехмерном пространстве. Сам Штейнер дал несколько иную, более сложную трактовку этой проблемы, пригодную для пространственного случая, но мы не приводим ее, так как на ее основе трудно получить доказательство существования решения. Вообще доказательство изопериметрического свойства сферы гораздо труднее, чем доказательство соответствующего свойства окружности; в достаточно полном и строгом изложении оно было дано позднее Г. А. Шварцем в работе, чтение которой довольно затруднительно. Свойство, о котором мы говорим, выражается в виде неравенства

$$36\pi V^2 \leqslant A^3,$$

где A — площадь замкнутой поверхости, V — охватываемый ею объем; равенство осуществляется лишь для сферы.

*§ 9. Экстремальные проблемы с граничными условиями. Связь между проблемой Штейнера и изопериметрической проблемой

Решение экстремальных проблем принимает своеобразные черты, если область значений переменного подчинена тем или иным граничным условиям. Теорема Вейерштрасса (утверждающая, что в компактной области непрерывная функция принимает наибольшее и наименьшее значения) не исключает возможности того, что эти экстремальные значения достигаются на границе области. В качестве простого, почти тривиального примера может служить функция $u = x$. Если x не подчинено никаким ограничениям и может изменяться от $-\infty$ до $+\infty$, то область B независимого переменного есть вся действительная ось; отсюда легко понять, что функция $u = x$ нигде не принимает ни наибольшего, ни наименьшего значения. Но если область B ограничена, например, неравенством $0 \leqslant x \leqslant 1$, то налицо имеется и наибольшее значение 1, достигаемое на правом конце промежутка, и наименьшее значение 0, достигаемое на левом. Но этим экстремальным значениям не соответствует «вершина» или «впадина» графика рассматриваемой функции. Иначе говоря, эти экстремумы осуществляются относительно не «двусторонней» окрестности; оставаясь на концах промежутка, они смещаются при расширении рассматриваемого промежутка. Если речь идет о настоящей «вершине» или «впадине» кривой, то экстремальный характер относится к полной окрестности рассматриваемой точки; небольшие сдвиги границы промежутка никак не влияют на экстремум. Такого рода экстремум сохраняется даже при свободном изменении переменного во всей области B или по крайней мере в некоторой достаточно малой окрестности точки. При самых разнообразных обстоятельствах поучительно уяснить себе

различие между «свободными» и «граничными» экстремумами. В случае функции одной переменной это различие, правда, стоит в тесной связи со свойствами монотонности или немонотонности функции и потому не приводит к каким-нибудь особенно интересным замечаниям. Но стоит остановиться несколько внимательнее на условиях достижения экстремума на границе области изменения в случае функций многих переменных.

Рассмотрим, например, проблему Шварца, касающуюся треугольника. Область изменения трех независимых переменных состоит здесь из троек точек P, Q, R, лежащих соответственно на сторонах треугольника ABC. Решение проблемы носит альтернативный характер: или минимум достигается при условии, что каждая из трех независимо движущихся точек P, Q, R находится внутри соответствующей стороны треугольника (и тогда задача решается высотным треугольником), или же минимум достигается «на границе», когда какие-то две из точек P, Q, R совпадают с общим концом двух смежных сторон (и тогда минимальный «треугольник» есть не что иное, как дважды считаемая высота данного треугольника). Характер решения — тот или иной, смотря по тому, которая из возможностей имеет место.

В проблеме Штейнера, относящейся к трем «деревням», область изменения точки P есть вся плоскость, причем данные три точки A, B, C могут считаться граничными. И в этом случае возникают две возможности, дающие решение существенно различного характера: или минимум достигается внутри треугольника ABC (и тогда около точки P возникают три равных угла), или он достигается в одной из вершин — граничных точек области изменения. Подобные альтернативы имеют место и для дополнительной проблемы.

Рассмотрим, наконец, в качестве последнего примера изопериметрическую проблему с добавочными граничными условиями. Мы установим при этом замечательную связь между изопериметрической проблемой и проблемой Штейнера и, помимо того, повстречаемся с простейшим примером экстремальной проблемы нового типа. В исходной изопериметрической проблеме замкнутая кривая данной длины, играющая роль независимого переменного, может быть свободно деформируема, как угодно отклоняясь от окружности, и любая получаемая кривая является «допустимой»; таким образом, окружность дает настоящий свободный минимум. Видоизмененная проблема содержит дополнительное требование: допустимые кривые C должны заключать внутри себя данные точки P, Q, R (или должны проходить через них); как и раньше, площадь A считается заданной, и предлагается минимизировать длину L. В этом примере мы имеем «граничное» условие в настоящем смысле слова.

Ясно, что при достаточно большом значении A три точки P, Q, R не оказывают на решение проблемы никакого влияния. В самом деле, если

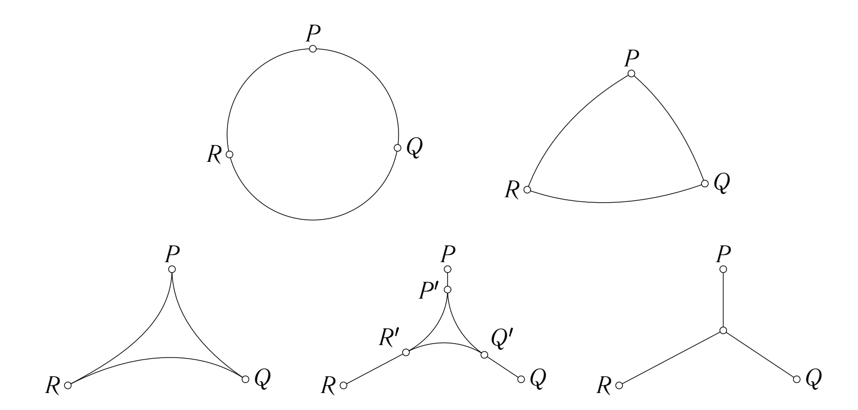


Рис. 231–235. Изопериметрические фигуры, в пределе дающие решение проблемы Штейнера

только A больше (или равно) площади круга, описанного около треугольника PQR, решение дается просто-напросто окружностью, охватывающей эти точки. Но что получается в противном случае? Укажем только результаты, оставляя в стороне детали доказательства, впрочем, вполне элементарного. Итак, постараемся охарактеризовать решение проблемы, предполагая, что данное числовое значение A постепенно становится меньше и, наконец, обращается в нуль. Как только A делается меньше, чем площадь описанного круга, изопериметрическая окружность превращается в три круговые дуги одного и того же радиуса, образующие выпуклый треугольник с вершинами P, Q, R (рис. 232). Этот треугольник и дает решение проблемы; он определяется полностью числовым значением A. При убывании A радиус дуг увеличивается, и дуги выпрямляются; когда A становится равным площади треугольника PQR, этот самый треугольник и дает решение. Если A становится еще меньше, то снова получаются треугольники, составленные из круговых дуг одного и того же радиуса, но с выпуклостью, обращенной внутрь треугольника, с вершинами — или, лучше сказать, «рожками» — в точках P, Q, R (рис. 233). При дальнейшем убывании A наступит момент, когда две круговые дуги, смыкающиеся у одной из данных точек, например R, станут касательными друг к другу. Еще далее, треугольники указанного типа уже перестанут быть возможными, и тогда обнаруживается новое явление: решение, как и перед тем, дается вогнутым треугольником, составленным из круговых дуг, но один из «рожков» R' отделяется от точки R, и решение тогда состоит из кругового треугольника PQR' с добавлением «дважды считаемого» (от R' к R и обратно) прямолинейного отрезка RR'. Этот отрезок касается двух круговых дуг, смыкающихся в

точке R'. Когда A убывает еще дальше, «рожки» отделяются и у прочих вершин. При достаточно малых положительных значениях A мы будем иметь равносторонний треугольник, составленный из трех круговых дуг одного и того же радиуса, касающихся друг друга в вершинах P', Q', R', с добавлением трех «дважды считаемых» отрезков $P'P$, $Q'Q$, $R'R$ (рис. 234). Наконец, при обращении A в нуль названный треугольник обращается в точку, и мы получаем решение проблемы Штейнера, которая, таким образом, оказывается предельным случаем обобщенной (указанным выше способом) изопериметрической проблемы.

Если P, Q, R образуют тупоугольный треугольник с углом в 120° или больше, то при стремлении A к нулю в пределе также получается решение проблемы Штейнера, так как круговые дуги в конце концов сливаются со сторонами тупого угла. Аналогичным образом, путем предельного перехода от изопериметрической проблемы, могут быть получены и решения обобщенной проблемы Штейнера (см. рис. 216–218 на стр. 382).

§ 10. Вариационное исчисление

1. Введение. Изопериметрическая проблема представляет собой, пожалуй, самый старый пример обширного класса важных проблем, к которым было привлечено общее внимание в 1696 г. Иоганном Бернулли. В «Acta Eruditorum», выдающемся научном журнале той эпохи, он поставил следующую проблему «о брахистохроне». Материальная частица скользит без трения по некоторой кривой, соединяющей выше расположенную точку A с ниже расположенной точкой B. Предполагая, что на частицу не действуют никакие силы, кроме силы тяжести, требуется установить, какова должна быть кривая AB, чтобы время, нужное для спуска от A к B, было наименьшим. Легко понять, что для спуска частицы от A к B необходимо то или иное время в зависимости от выбора пути. Прямолинейный отрезок никоим образом не обеспечивает наименьшего времени; то же приходится сказать о круговых дугах и других элементарных кривых. Бернулли объявил, что он обладает замечательным решением поставленной задачи, которого, однако, не хочет пока публиковать, имея в виду побудить крупнейших математиков своего времени приложить свое искусство к математическим задачам нового типа. В частности, он вызвал на состязание своего старшего брата Якоба, с которым был тогда в резко враждебных отношениях и открыто именовал невеждой. Своеобразие задачи о брахистохроне вскоре действительно было оценено математическим миром. В проблемах, исследованных до того времени с помощью дифференциального исчисления, подлежащая минимизации величина зависела от одной или нескольких (в конечном числе) числовых переменных; в этой же задаче

рассматриваемая величина — время спуска — зависит от всей кривой в целом, чем и обусловливается существенное различие; именно по указанной причине задача о брахистохроне не могла быть решена ни методом дифференциального исчисления, ни каким-либо другим, известным в те времена приемом.

Новизна поставленной проблемы (по-видимому, то обстоятельство, что доказательство изопериметрического свойства круга представляет собой вопрос той же природы, не было тогда еще осознано) подействовала на современников Бернулли, в особенности когда выяснилось, что решением задачи является циклоида — как раз незадолго до того открытая кривая. (Напомним определение циклоиды: так называют траекторию движения точки, находящейся на окружности, которая катится без скольжения по прямой линии — рис. 236. Эта кривая уже раньше была поставлена в связь с некоторыми интересными задачами механического содержания, в частности, с конструированием идеального маятника.) Гюйгенс установил, что если тяжелая частица (точка) совершает (без трения, под влиянием силы тяжести) колебательное движение по циклоиде, расположенной в вертикальной плоскости, то период колебания не зависит от амплитуды (размаха). Напротив, на круговой дуге, представляющей собой траекторию движения обыкновенного маятника, такого рода независимость имеет лишь приближенный характер, и в этом обстоятельстве усматривалась непригодность круговой дуги при конструировании точных часов. Циклоиде было присвоено, в связи с указанным обстоятельством, наименование таутохроны, но теперь она стала именоваться также и брахистохроной[1].

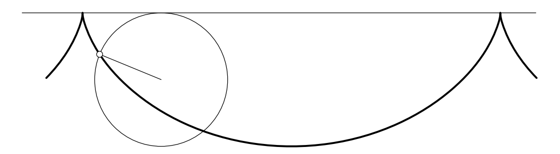


Рис. 236. Циклоида

2. Вариационное исчисление. Принцип Ферма в оптике. Среди различных методов, с помощью которых решение брахистохронной проблемы было найдено братьями Бернулли и другими учеными, мы выберем и изложим здесь один из самых ранних в историческом смысле. Первые предложенные методы носили более или менее специальный характер, будучи более приспособлены к специфическим задачам. Но очень скоро Эйлер и Лагранж (1736–1813) разработали более общие методы для решения экстремальных проблем, в которых независимым

[1] *Таутохрона* — от греч. ταυτο (равно), χρουος (время); *брахистохрона* — от греч. βραχυς (короткий), χρουος. — *Прим. ред.*

элементом является не одна или несколько (в конечном числе) числовых переменных, а кривая или функция в целом, или даже система кривых (функций). Новый метод решения подобного рода проблем получил название *вариационного исчисления*.

Дать здесь изложение этой ветви математики в ее техническом аспекте или же проанализировать сколько-нибудь глубоко отдельные относящиеся сюда проблемы не представляется возможным. Вариационное исчисление имеет множество применений в физических теориях. Было замечено с давних пор, что явления природы часто следуют тем или иным экстремальным принципам. Как мы уже видели, Герон Александрийский усмотрел, что отражение светового луча плоским зеркалом хорошо описывается на основе принципа минимума. Ферма — уже в XVII столетии — сделал следующий шаг, заметив, что и закон преломления света также прекрасно выражается в терминах минимального принципа. Отлично известно, что при переходе светового луча из одной однородной среды в другую путь его изменяет направление. Так, световой луч, идущий из точки P (рис. 237) в верхней среде, где скорость равна v, в точку R в нижней среде, где скорость есть w, совершит ломаный путь PQR. Снеллиус (1591–1626) сформулировал найденный им эмпирическим путем закон, согласно которому путь состоит из двух прямолинейных отрезков PQ и QR, образующих с нормалью углы α и α', причем $\frac{\sin\alpha}{\sin\alpha'} = \frac{v}{w}$. С помощью дифференциального исчисления Ферма установил, что этот путь как раз обладает тем свойством, что время, нужное для прохода луча из P в R, минимально, т. е. меньше, чем понадобилось бы при прохождении по любому иному пути. Таким образом, спустя шестнадцать столетий героновзакон отражения света был дополнен подобным ему и столь же важным законом преломления.

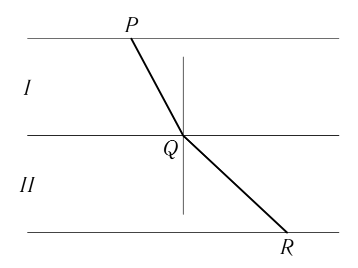


Рис. 237. Преломление светового луча

Ферма обобщил формулировку этого закона, распространяя его на случай кривых поверхностей раздела между двумя средами, каковы, например, сферические поверхности линз. Оказывается, что и в этом случае световой луч следует пути, обладающему тем свойством, что время, нужное для его прохождения, меньше, чем понадобилось бы при выборе любого другого пути. Наконец, Ферма рассмотрел и случай произвольной оптической системы, в которой скорость света меняется по определенному закону от точки к точке, например так, как это происходит в атмосфере. Он разделил непрерывную неоднородную среду

на тонкие слои, в каждом из которых скорость света приблизительно постоянна, и представил себе новую, воображаемую среду, в которой скорость света действительно постоянна в пределах каждого слоя. При таких условиях можно было применять прежний принцип при переходе от каждого слоя к следующему. Затем, допуская, что толщина каждого слоя стремится к нулю, он получил *общий принцип геометрической оптики* (известный ныне под именем принципа Ферма): в неоднородной среде световой луч, идущий от одной точки к другой, следует по такому пути, что время, нужное для его прохождения, меньше, чем понадобилось бы при прохождении любого иного пути. Этот принцип оказался в высшей степени полезным не только теоретически, но и практически. В геометрической оптике, оперируя техническим аппаратом вариационного исчисления, пользуются этим принципом как основным орудием при расчетах систем линз.

Минимальные принципы стали затем господствующими и в других областях физики. Так, было замечено, что устойчивое равновесие механической системы бывает достигнуто при таком расположении, при котором «потенциальная энергия» минимальна. Рассмотрим, например, свободно изгибаемую однородную цепь, подвешенную за два ее конца и предоставленную действию силы тяжести. Тогда цепь займет именно такое положение, при котором потенциальная энергия ее будет наименьшей. В указанном примере потенциальная энергия зависит от высоты центра тяжести относительно некоторой постоянной оси. Кривая, образованная свободно подвешенной цепью, называется *цепной линией* и по внешнему виду несколько напоминает параболу.

Не только закон равновесия, но и законы движения подчиняются экстремальным принципам. Отчетливые представления об этих принципах впервые возникли у Эйлера, тогда как люди, склонные к спекулятивным размышлениям философского и мистического характера, как, например, Мопертюи (1698–1759), не были способны дать точные математические формулировки и ограничивались смутными высказываниями по поводу «божественного регулирования физических явлений общими принципами наивысшего совершенства». Эйлеровы вариационные принципы в области физики, вновь открытые и обобщенные ирландским математиком Гамильтоном (1805–1865), стали впоследствии могущественнейшим орудием в таких областях, как механика, оптика, электродинамика, самые разнообразные технические науки. Физические теории недавнего происхождения — теория относительности и квантовая теория — полны примеров, обнаруживающих значение методов вариационного исчисления.

3. Решение задачи о брахистохроне, принадлежащее Якобу Бернулли. Ранний метод, примененный к решению проблемы о брахистохроне Якобом Бернулли, может быть изложен с применени-

ем сравнительно скромных математических средств. Возьмем в качестве исходного тот известный из механики факт, что материальная частица, начинающая свой путь в точке A с нулевой скоростью и затем скользящая вниз по произвольной кривой C, приходит в некоторую точку P со скоростью, пропорциональной величине $\sqrt{h}$, где h есть отсчитываемое по вертикали расстояние точки P от точки A; иначе говоря, мы имеем зависимость $v = c\sqrt{h}$, где c — постоянный коэффициент. Подвергнем рассматриваемую задачу легкому видоизменению. Разобьем мысленно пространство на множество горизонтальных слоев, каждый толщиной d, и предположим на минуту, что скорость нашей частицы меняется не непрерывно, а небольшими скачками — при переходе от слоя к слою; именно в первом слое, прилежащем непосредственно к точке A, скорость равна $c\sqrt{d}$, во втором $c\sqrt{2d}$, наконец в n-м $c\sqrt{nd} = c\sqrt{h}$, где h — расстояние P от A, отсчитываемое по вертикали (рис. 238). При такой постановке задачи мы имеем дело с конечным числом переменных. В пределах каждого слоя путь частицы должен быть прямолинейным. Вопрос о существовании экстремума не возникает; решение должно даваться ломаной линией; нужно только определить ее углы при вершинах. Согласно минимальному принципу простого преломления, в каждой паре соседних слоев движение от P к R через Q таково, что при фиксированных P и R точка Q соответствует наименьшему времени пути. Отсюда вытекает следующий «закон преломления»:

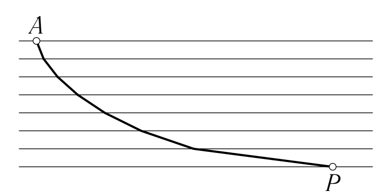


Рис. 238. К проблеме брахистохроны

$$\frac{\sin\alpha}{\sqrt{nd}} = \frac{\sin\alpha'}{\sqrt{(n+1)d}}.$$

Повторное применение этого рассуждения приводит к цепи равенств

$$\frac{\sin\alpha_1}{\sqrt{d}} = \frac{\sin\alpha_2}{\sqrt{2d}} = \ldots, \qquad (1)$$

где α_n обозначает угол между направлением пути в n-м слое и вертикалью.

Затем Бернулли предполагает, что толщина слоев d, неограниченно уменьшаясь, стремится к нулю, причем ломаная траектория, решающая приближенную проблему, в пределе переходит в искомую кривую, решающую основную проблему. При этом предельном переходе равенства (1) сохраняются, и потому Бернулли делает заключение: если α обозначает угол, который в произвольной точке P кривой C траектория брахистохронного движения делает с вертикалью, а h — расстояние от A до P, рассчитываемое по вертикали, то выражение $\dfrac{\sin\alpha}{\sqrt{h}}$ должно сохранять

постоянное значение во всех точках P кривой C. Легко показать, что указанное свойство характеризует циклоиду.

Бернуллиево «доказательство» представляет собой типичный пример остроумного и плодотворного математического рассуждения, которое в то же время нельзя назвать безукоризненно строгим. В нем содержится несколько неявно принятых допущений, оправдание которых было бы сложнее и пространнее, чем само рассуждение. Так, с одной стороны, не доказывается само существование решения C, с другой — постулируется без достаточных математических оснований, что решение приближенной проблемы является приближенным решением основной проблемы. Вопрос о внутренней ценности такого рода эвристических (наводящих) построений заслуживает внимательного рассмотрения, но завел бы нас слишком далеко в сторону.

4. Геодезические линии на сфере. Минимаксы. Во введении к этой главе была упомянута проблема нахождения «геодезических линий» — кратчайших дуг, соединяющих две данные точки на некоторой поверхности. На сфере, как показывается в элементарной геометрии, такими линиями являются дуги больших кругов. Пусть P и Q — две точки на сфере (не являющиеся диаметрально противоположными) и c — меньшая из двух дуг большого круга, проходящего через P и Q. Тогда возникает вопрос: чем же является другая, бо́льшая из двух дуг c' того же круга. Конечно, минимума расстояния между точками P и Q она не дает, но не дает и максимума, так как легко понять, что можно провести на сфере сколь угодно длинные линии, соединяющие две данные точки. Оказывается, что по отношению к рассматриваемой проблеме дуга c' представляет собой минимакс, «седловую точку». Вообразим произвольную переменную точку S на сфере и поставим задачей найти кратчайший путь от P к Q, проходящий через S. Конечно, минимум расстояния в такой постановке проблемы дается «ломаной» дугой, состоящей из двух дуг больших кругов PS и SQ. А затем постараемся найти такое положение точки S, при котором наименьшее расстояние PSQ было бы максимальным. Тогда получаем следующее решение вопроса: точка S должна быть такова, чтобы ломаная PSQ была более длинной дугой c' большого круга PQ. Можно видоизменить проблему, сначала спрашивая себя о кратчайшем пути на сфере от точки P к точке S, проходящем через n наперед заданных точек S_1, S_2, ..., S_n, и затем определяя точки S_1, S_2, ..., S_n таким образом, чтобы минимальная длина была насколь-

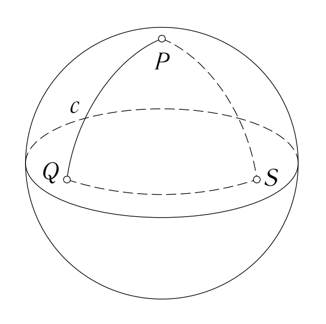


Рис. 239. Геодезические линии на сфере

ко возможно большой. Решением такой задачи служит путь по большому кругу, проходящему через P к Q, но обвивающийся вокруг сферы таким образом, чтобы пройти через точки, диаметрально противоположные P и Q, ровно n раз.

Эта минимаксная проблема является типичным примером для обширного класса вопросов из области вариационного исчисления, с полным успехом изученных в последнее время с помощью методов, предложенных Морсом и другими авторами.

§ 11. Экспериментальные решения задач на минимум. Опыты с мыльными пленками

1. Введение. Обыкновенно бывает очень трудно, а иногда даже невозможно, решить вариационную проблему явно с помощью формул или геометрических построений, включающих простые, известные элементы. Вместо того часто удовлетворяются одним лишь доказательством существования решения при тех или иных условиях и затем исследуют его свойства. Во многих случаях, если доказательство существования оказывается более или менее затруднительным, бывает полезно реализовать математические условия проблемы посредством соответствующих физических приспособлений, рассматривая таким образом математическую проблему как эквивалентную некоторой физической задаче. Само физическое явление в таких случаях предоставляет решение математической проблемы. Само собой разумеется, что подобного рода процедуру следует трактовать не как полноценное математическое доказательство, а только как «наводящую» (эвристическую): в самом деле, при этом остается открытым вопрос о том, является ли математическая интерпретация строго адекватной физическому явлению или же дает лишь несовершенное отображение реальной действительности. Иногда относящиеся сюда эксперименты, хотя бы они были воображаемыми, бывают способны воздействовать убеждающе даже на математиков. В прошлом столетии ряд фундаментальных теорем из области теории функций был открыт Риманом на основе придумывания простейших экспериментов, касающихся потока электричества в металлических листах.

В дальнейшем мы имеем в виду рассмотреть — на экспериментально-демонстративной основе — одну из более глубоких вариационных проблем. Речь идет о так называемой проблеме Плато. Плато (1801–1883), известный физик, по национальности бельгиец, занимался интересными опытами, имеющими ближайшее отношение к этой проблеме. Сама по себе проблема гораздо старше по возрасту и относится к эпохе возникновения вариационного исчисления. В простейшей формулировке содержание ее таково: найти поверхность наименьшей площади, ограничен-

ную данным замкнутым пространственным контуром. Мы рассмотрим также эксперименты, относящиеся к некоторым близким проблемам, и убедимся, что это позволит увидеть в новом свете как некоторые из приведенных выше экстремальных проблем, так и ряд экстремальных проблем нового типа.

2. Опыты с мыльными пленками. В математической постановке проблема Плато приводит к решению «дифференциального уравнения в частных производных» или же системы таких уравнений. Эйлер установил, что всякая «минимальная» поверхность, решающая эту проблему, если только не сводится к плоскости, непременно должна быть во всех своих точках «седлообразной» и что ее средняя кривизна всюду должна равняться нулю[1]. В течение последнего столетия решение было получено во множестве частных случаев, но существование решения в общем случае было доказано лишь недавно Дж. Дугласом и Т. Радо.

Опыты Плато непосредственно дают физические решения для самых разнообразных контуров. Если замкнутый контур, сделанный из проволоки, погрузить в жидкость со слабым поверхностным натяжением и затем вынуть оттуда, то увидим пленку, натянутую на контуре в форме минимальной поверхности с наименьшей площадью. (Предполагается, что можно пренебречь силой тяжести и другими силами, препятствующими стремлению пленки достигнуть устойчивого равновесия; последнее же наступает в том случае, если площадь пленки оказывается наименьшей, так как потенциальная энергия, возникающая вследствие поверхностного натяжения, при этом условии минимальна.) Вот хороший рецепт для получения такой жидкости: растворите 10 г чистого сухого олеата натрия в 500 г дистиллированной воды и затем смешайте 15 кубических единиц раствора с 11 кубическими единицами глицерина. Пленки, получаемые из указанной смеси на каркасах из латунной проволоки, сравнительно устойчивы. Сами каркасы не должны превышать 5–6 дюймов в диаметре.

С помощью пленок очень легко «решить» проблему Плато: достаточно придать проволочному каркасу нужную форму. Красивые модели поверхностей получаются на полигональных каркасах, образованных из последовательностей ребер правильных многогранников. В частности, любопытно погрузить в наш раствор каркас куба весь целиком. То-

[1] Средняя кривизна поверхности в точке P определяется следующим образом. Вообразим перпендикуляр к поверхности в точке P и все плоскости, через него проходящие. Эти плоскости пересекаются с данной поверхностью по кривым, которые в точке P имеют, вообще говоря, различную кривизну. Рассмотрим, в частности, кривые, обладающие наибольшей и наименьшей кривизной (соответствующие секущие плоскости, как можно доказать, перпендикулярны между собой). Полусумма этих двух кривизн и есть средняя кривизна поверхности в точке P.

гда получается система поверхностей, пересекающих друг друга под углом в 120°. (Если куб вынимать из раствора очень осторожно, то можно насчитать тринадцать почти плоских поверхностей.) Потом можно протыкать и уничтожать поверхности одну за другой, пока не останется только одна поверхность, ограниченная замкнутым полигональным контуром. Таким образом можно получить целый ряд прекрасных поверхностей. Тот же опыт можно проделать и с тетраэдром.

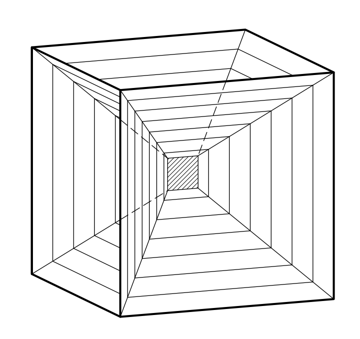


Рис. 240. На кубическом каркасе натянуто 13 почти плоских поверхностей

3. Новые опыты, относящиеся к проблеме Плато. Опыты с пленками не сводятся к демонстрации минимальной поверхности, натянутой на замкнутый контур (как у Плато); диапазон их гораздо шире. В последнее время проблема минимальных поверхностей была изучена не только для одного ограничивающего контура, но и для системы таких контуров; кроме того, было обращено внимание и на возможность образования минимальных поверхностей более сложной топологической структуры. Так, существуют односторонние минимальные поверхности и минимальные поверхности рода, отличного от нуля. Возникающие более общие проблемы порождают изумительное разнообразие геометрических явлений, которые могут быть продемонстрированы с помощью мыльных пленок. Заметим в связи с этим, что очень полезно проволочные каркасы делать гибкими и изучать изменение формы поверхности пленки под влиянием непрерывной деформации каркаса. Дадим описание некоторых опытов.

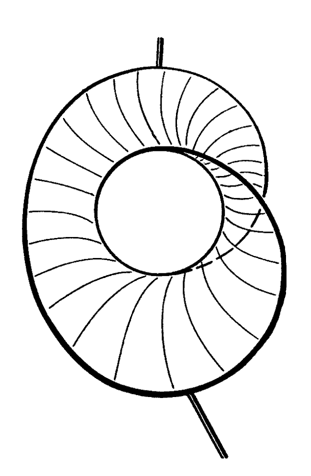


Рис. 241. Односторонняя поверхность (лента Мёбиуса)

1. Если граничный контур представляет собой окружность, то получается поверхность в виде кругового диска. Можно было бы ожидать, что при непрерывной деформации контура минимальная поверхность всегда будет сохранять тот же топологический характер. Но это неверно. Если изогнуть контур так, как показано на рис. 241, то вместо поверхности, топологически эквивалентной диску, получается односторонняя лента Мёбиуса. Обратно, можно производить

деформацию, исходя из контура, изображенного на чертеже, с натянутой на него пленкой в виде ленты Мёбиуса. Для осуществления непрерывной деформации следует припаять к каркасу рукоятки (см. тот же рисунок). В процессе обратной деформации наступает момент, когда внезапно топологический характер пленки меняется и возникает снова поверхность типа диска (рис. 242). Опять, обращая деформацию, мы вернемся к поверхности Мёбиуса. Замечательно, однако, то, что «мутация» дискообразной поверхности в поверхность типа Мёбиуса происходит на более поздней стадии деформации, чем при обратном процессе. Это показывает, что существует непрерывная цепь замкнутых пространственных контуров, для которых и поверхности типа Мёбиуса, и дискообразные поверхности устойчивы, т. е. доставляют относительные минимумы. Но если поверхность типа Мёбиуса обладает значительно меньшей площадью, чем другая, то эта последняя все же слишком неустойчива, чтобы существовать физически.

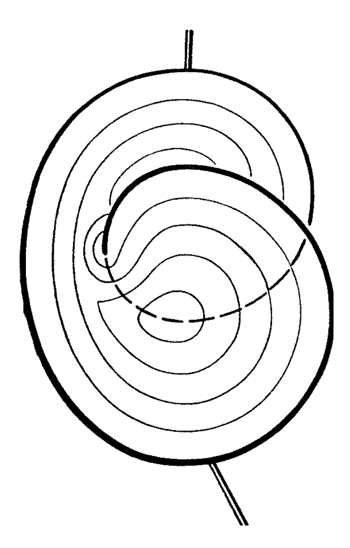


Рис. 242. Двусторонняя поверхность

2. Можно натянуть минимальную поверхность на систему контуров, состоящих из двух окружностей. Вынув каркас из раствора, мы получаем не одну поверхность, а структуру, состоящую из трех поверхностей, смыкающихся под углом в 120°; одна из них — обыкновенный круговой диск, плоскость которого параллельна плоскостям граничных окружностей (рис. 243). Уничтожая этот диск, мы получим, далее, классический катеноид (поверхность, образуемую вращением цепной линии, о которой шла речь на стр. 404, около прямой, перпендикулярной к оси симметрии). При раздвигании граничных контуров наступает момент, когда двусвязный катеноид лопается и превращается в два отдельных диска. Указанный процесс, конечно, необратим.

3. Еще один замечательный пример доставляется каркасом, изображенным на рис. 244–246; на этот каркас могут быть натянуты три различные минимальные поверхности. Одна из них (рис. 244) имеет род 1, тогда как две другие односвязны и в некотором смысле обладают свойством взаимной симметрии. Две последние поверхности имеют одну и ту же площадь, если только контур вполне симметричен. Но в противном случае только одна из поверхностей обеспечивает абсолютный минимум площади, тогда как другая — только относительный (мы предполагаем при этом, что минимум разыс-

кивается только по отношению к односвязным поверхностям). Возможность образования поверхности рода 1 обусловливается тем обстоятельством, что, допуская поверхности рода 1, можно получить поверхность меньшей площади, чем для какой бы то ни было односвязной поверхности. При деформации контура мы придем, если только деформация будет достаточно сильно выражена, к такому положению, когда указанное свойство будет уже утеряно: тогда поверхность рода 1 потеряет свою устойчивость и, внезапно разрываясь, превратится в односвязную поверхность одного из двух типов, изображенных на рис. 245 и 246. Если, с другой стороны, мы станем исходить из поверхности одного из этих двух типов, — например, изображенного на рис. 246, то возможно деформировать контур таким образом, что другой тип (см. рис. 245) станет гораздо более устойчивым. Следствием этого явится тот факт, что в определенный момент произойдет внезапный переход от одного типа к другому. Медленно обращая всю деформацию в обратном направлении, вернем контур снова к исходному положению, но уже с иной натянутой на нем минимальной поверхностью. Можно снова повторить весь процесс в обратном направлении; таким образом, можно многократно повторять переход от одного типа поверхности к другому. Оперируя контуром надлежащим образом, удается

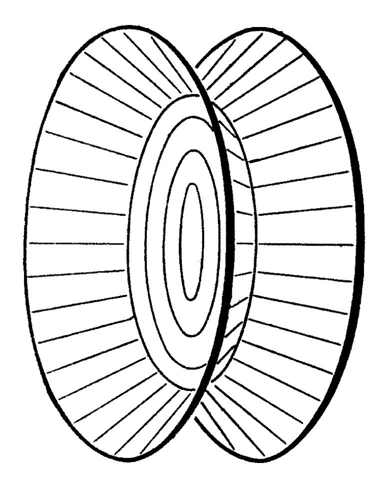


Рис. 243. Система трех поверхностей

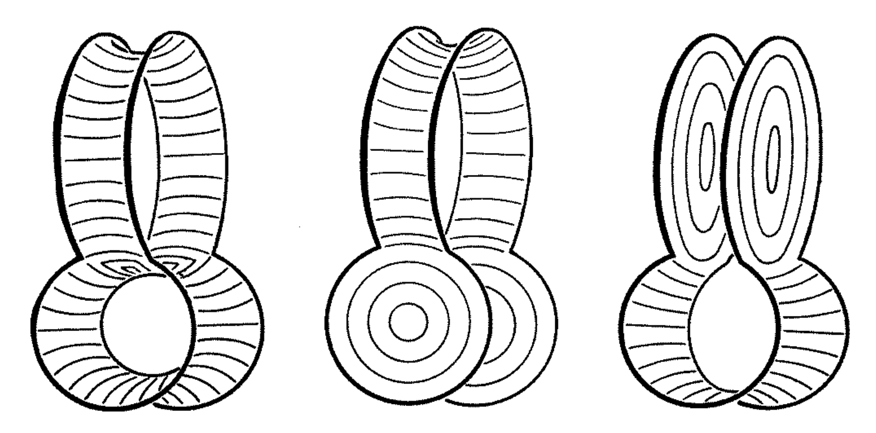


Рис. 244–246. На каркасе натянуты три различные поверхности родов 0 и 1

также трансформировать одну из односвязных поверхностей в поверхность рода 1. Для этой цели нужно сблизить между собой те части

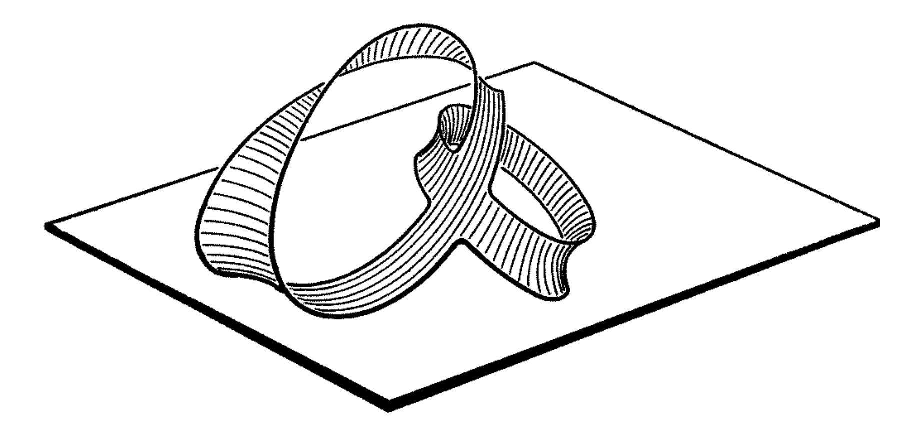


Рис. 247. Односторонняя минимальная поверхность более сложной топологической структуры, натянутая на простой замкнутый контур

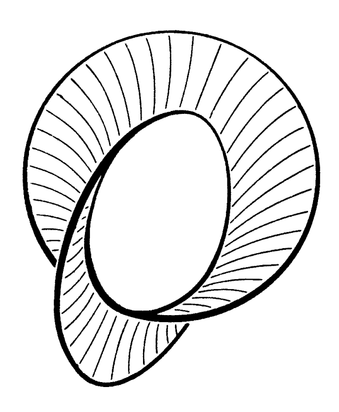


Рис. 248. Поверхность, натянутая на два зацепленных круга

контура, на которые натянуты дискообразные части самой поверхности — с таким расчетом, чтобы поверхность рода 1 стала гораздо более устойчивой. Иногда в процессе выполнения описанной выше операции с контуром возникают промежуточные пленочные поверхности: их нужно уничтожать, чтобы получилась поверхность рода 1.

Этот пример показывает не только возможность решения проблемы Плато различными поверхностями одного и того же топологического типа, но и поверхностями иного типа, причем на одном и том же контуре; кроме того, он снова иллюстрирует прерывный переход от одного решения к другому, в то время как граничные условия проблемы меняются непрерывно. Нетрудно построить более сложные модели в таком же роде и подвергнуть их экспериментальному исследованию.

Интересное явление — возникновение минимальных поверхностей, ограниченных двумя или большим числом взаимно зацепленных замкнутых контуров. В случае двух круговых контуров получается поверхность, изображенная на рис. 248. Если в этом примере плоскости кругов взаимно перпендикулярны и прямая их пересечения есть общий диаметр

двух кругов, то существуют две диаметрально противоположные формы минимальной поверхности с одинаковыми площадями. Представим себе теперь, что два круга постепенно изменяют свое взаимное положение; тогда и форма минимальной поверхности будет меняться непрерывно, хотя при каждом положении кругов только для одной из поверхностей осуществляется абсолютный минимум, для другой же — только относительный. При некоторых положениях поверхность относительного минимума вдруг разрывается и заменяется поверхностью абсолютного минимума. Обе минимальные поверхности в этом примере — одного и того же топологического типа (как и поверхности на рис. 245 и 246).

4. Экспериментальные решения других математических проблем. Благодаря действию поверхностного натяжения жидкая пленка только при том условии может находиться в состоянии устойчивого равновесия, если площадь образуемой поверхности минимальна. Это обстоятельство является неисчерпаемым источником экспериментов серьезной математической ценности. Если некоторые части границы пленки могут свободно перемещаться по заданным поверхностям (например, плоскостям), то на этих частях границы пленка будет стоять перпендикулярно к заданной поверхности.

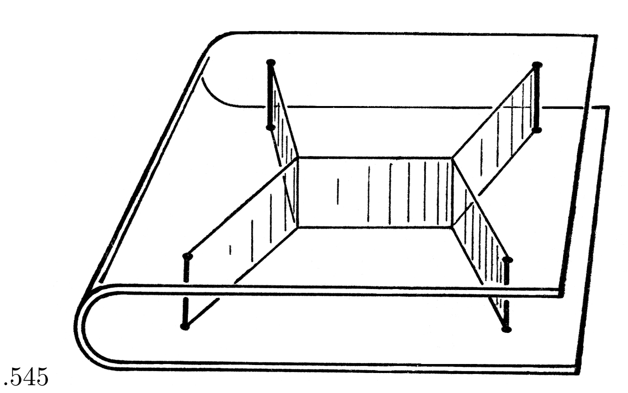


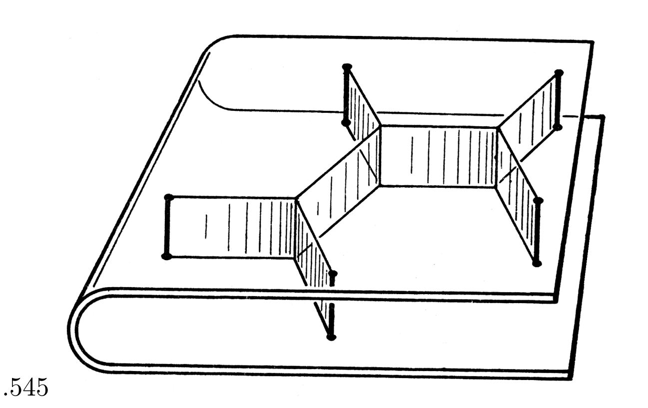
Рис. 249. Демонстрация кратчайшей системы путей меж-ду 4 точками

Мы можем использовать последнее отмеченное обстоятельство для наглядного решения проблемы Штейнера и ее обобщений (см. § 5). Пусть две параллельно расположенные стеклянные поверхности (или гладкие плитки) соединены тремя или бóльшим числом перпендикулярно стоящих стержней. Если погрузить всю такого рода систему в мыльный раствор, затем вынуть, то пленка образует между плоскими поверхностями ряд вертикальных полос, связывающих между собой стержни. Проекция этих полос на горизонтальные плоскости есть не что иное, как решение проблемы Штейнера, рассмотренной на стр. 376–377.

Если две плоские поверхности не параллельны или стержни к ним не перпендикулярны, или сами поверхности не являются плоскими, то кривые, по которым пленки пересекаются с поверхностями, не будучи пря-

мыми линиями, смогут иллюстрировать решение новых вариационных проблем.

Появление кривых, по которым смыкаются под углами в 120° различные минимальные поверхности, может рассматриваться как пространственное обобщение явлений, связанных с проблемой Штейнера. Это становится вполне ясным, если мы соединим, например, две точки A, B тремя различными пространственными кривыми и затем погрузим полученную (жестко укрепленную) систему в мыльный раствор. Предположим для определенности, что одна из трех кривых есть прямолинейный отрезок, две другие — взаимно конгруэнтные круговые дуги. То, что получается, изображено на рис. 251. Если плоскости дуг образуют между собой угол меньше 120°, мы получим решение минимальной проблемы в виде трех поверхностей, смыкающихся под



.545

Рис. 250. Кратчайшая система путей между 5 точками

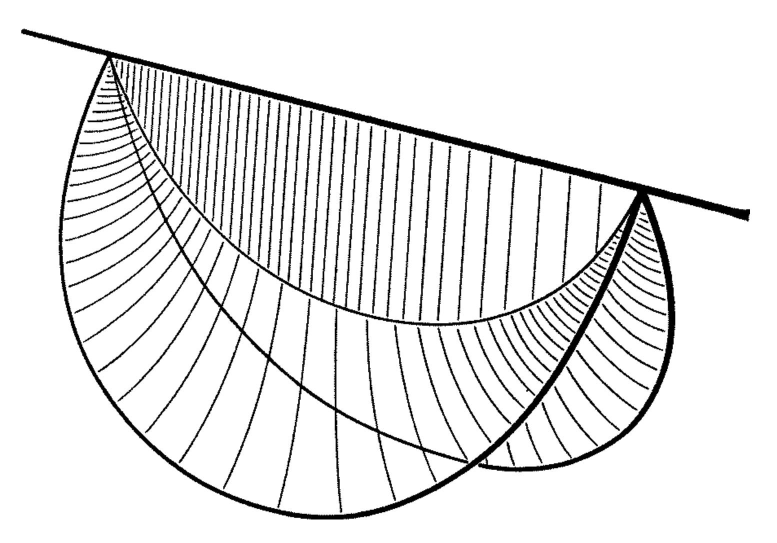


Рис. 251. Три пересекающиеся под углом и 120° поверхности, натянутые на три проволоки, соединяющие две точки

углами в 120°, но если станем поворачивать плоскости дуг, увеличивая заключенный между ними угол, то это решение в результате непрерывного изменения перейдет, наконец, в два плоских круговых сегмента.

Допустим теперь, что точки A и B соединены более сложными кривыми. В качестве примера возьмем три ломаные, состоящие каждая из трех ребер одного и того же куба и соединяющие диагонально противоположные
вершины: тогда получатся три конгруэнтные минимальные поверхности, пересекающиеся по диагонали куба. (Мы получили бы ту же систему поверхностей из системы, изображенной на рис. 240, уничтожая пленки, прилежащие к трем надлежащим образом выбранным ребрам.) Если станем деформировать ломаные линии, соединяющие A и B, то линия взаимного смыкания поверхностей искривится, но углы неизменно останутся те же — в 120° (рис. 252).

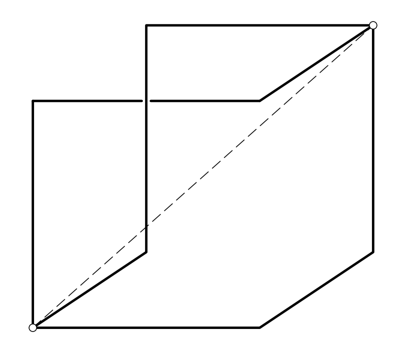
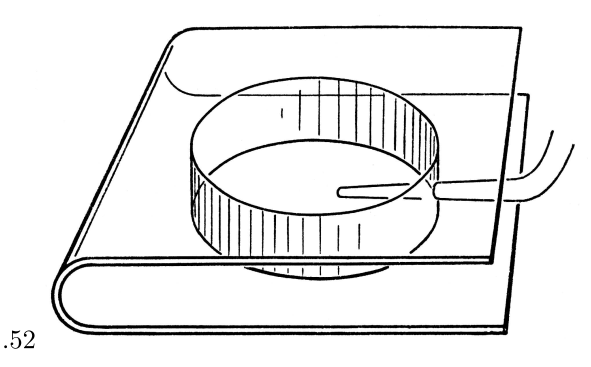


Рис. 252. Три ломаные линии, соединяющие две точки

Все явления, связанные со смыканием трех минимальных поверхностей по одной кривой, в основном одной и той же природы: они представляют собой обобщение плоской проблемы о соединении системы n данных точек кратчайшей системой линий.

Наконец, добавим несколько слов о мыльных пузырях. Сферический мыльный пузырь показывает, что среди всех замкнутых поверхностей, охватывающих один и тот же объем (определенный запасом заключенного в нем воздуха), именно сфера имеет наименьшую поверхность. Если мы рассмотрим пузыри данного объема, стремящиеся сократить свою поверхность, но подчиненные некоторым дополнительным условиям, то убедимся, что получаться будут уже не обязательно сферы, а, вообще говоря, поверхности постоянной средней кривизны, частными примерами которых являются сферы и круговые цилиндры.



.52

Рис. 253. Доказательство изопериметрического свойства круга

Предположим, например, что пузырь заключен между двумя параллельными стеклами или плитками, предварительно смоченными мыльным раствором. Прикоснувшись к одной из плоскостей, пузырь внезапно принимает форму полусферы, если же происходит соприкосновение также и с другой плоскостью, он сразу превращается в круговой цилиндр, тем самым чрезвычайно наглядно демонстрируя изопериметрическое свойство круга. Все дело, конечно, в том, что мыльная пленка располагается перпендикулярно к ограничивающим поверхностям. Помещая мыльные пузыри между двумя плоскостями, которые соединены между собой стержнями, мы имеем возможность проиллюстрировать проблемы, разобранные на стр. 381.

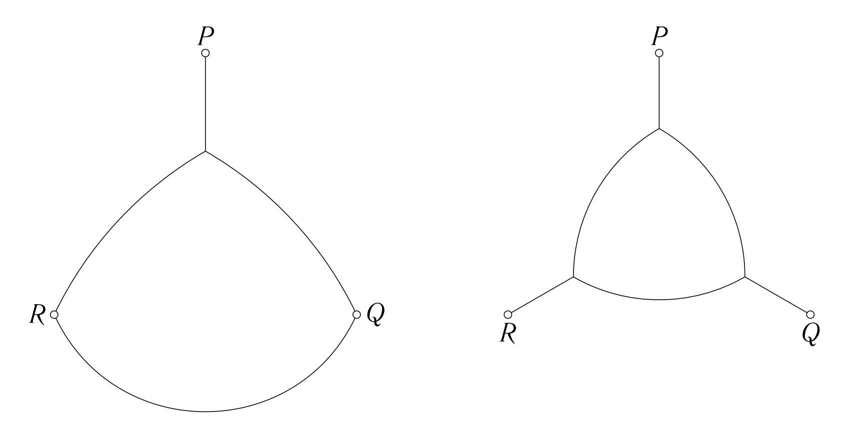
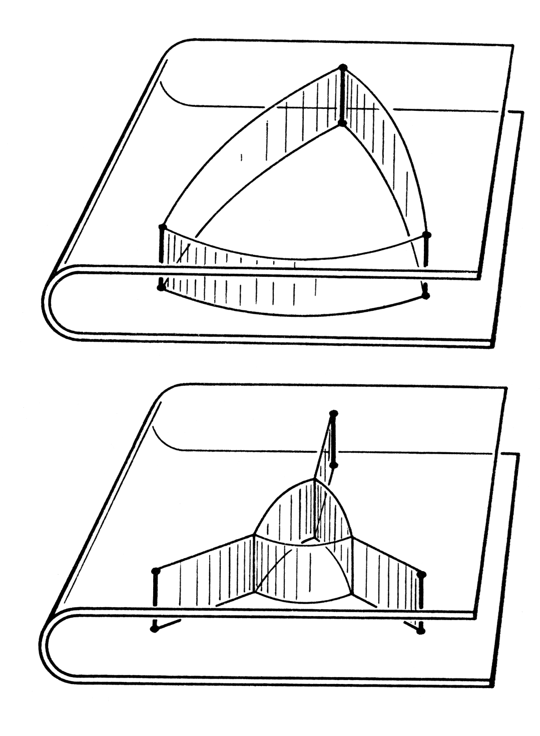


Рис. 254–255. Изопериметрические фигуры с граничными условиями

Можно еще рассмотреть, как изменяется решение изопериметрической проблемы при увеличении или уменьшении объема воздуха внутри пузыря. При этом следует воспользоваться тоненькой трубочкой или соломинкой. Однако, высасывая воздух, мы не получим тех фигур (см. стр. 400), которые состоят из касающихся друг друга круговых дуг. При уменьшении объема воздуха внутри пузыря углы в треугольнике из круговых дуг, однако, не станут (теоретически) меньшими, чем 120°: мы получим такие фигуры, какие изображены на рис. 254 и 255, причем при неограниченном уменьшении площади, заключенной внутри, в пределе получатся те же три сегмента, с которыми мы встретились и раньше (рис. 235). С математической точки зрения объяснение отмеченному различию заключается в том, что отрезок, связывающий пузырь с каким-нибудь стержнем, начиная с момента отделения пузыря от этого стержня, не должен считаться дважды. Соответствующие опыты иллюстрируются рис. 256 и 257.

Упражнение. Разберите математическую проблему, соответствующую



270pt

Рис. 256–257. Демонстрация изопериметрических свойств фигур с помощью мыльных пленок

следующим условиям: найти треугольник, составленный из круговых дуг и имеющий данную площадь, по условию, чтобы сумма его периметра и трех отрезков, соединяющих вершины с тремя данными точками, была минимальной.

Помещая мыльный пузырь внутрь кубического проволочного каркаса, в случае если объем пузыря окажется больше, чем объем куба, мы получим поверхности постоянной средней кривизны с квадратными основаниями. Высасывая воздух из пузыря через соломинку, будем иметь целую цепь красивых структур, приводящих, в конце концов, к такой, какая изображена на рис. 258. Явления устойчивости и переход от одного состояния равновесия к другому порождают эксперименты, которые в математическом отношении нельзя не назвать весьма поучительными. Таким образом, возникает наглядная иллюстрация к теории стационарных значений; непрерывная цепь переходов от одного состояния равновесия к другому может быть выбрана таким образом, что в ее состав войдет состояние неустойчивого равновесия, все же являющееся «стационарным состоянием».

Рассмотрим в качестве примера кубическую структуру на рис. 240. Мы видим здесь нарушение симметрии: в центре куба имеется вертикальная площадка, смыкающаяся с двенадцатью поверхностями, идущими от ребер куба. Но тогда, как нетрудно понять, должно существовать еще по меньшей мере два положения равновесия: одно с вертикальной (иначе расположенной) и другое с горизонтальной площадкой в центре. Чтобы на самом деле реализовать переход от одного положения равновесия к другому, нужно дуть через соломинку на ребра центральной площадки: при этом удается центральную площадку превратить в точку — центр куба, но полученное таким образом состояние равновесия не будет устойчивым и немедленно же перейдет в иное устойчивое состояние, причем центральная площадка снова возникает, хотя и повернувшись на 90°.

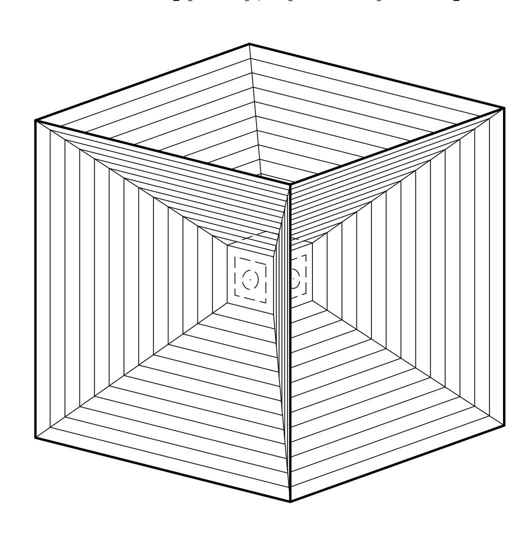


Рис. 258. Пленки на кубическом каркасе

Подобный же эксперимент можно произвести и с мыльной пленкой, демонстрирующей решение проблемы Штейнера для случая четырех точек, помещенных в вершинах квадрата (рис. 219, 220).

Если бы мы пожелали получить решение только что рассмотренных проблем как предельный случай цепи изопериметрических проблем, например, если бы мы хотели получить рис. 240 из рис. 258, нужно было бы понемногу высасывать воздух из центрального пузыря. Структура, изображенная на рис. 258, строго симметрична, и в пределе, когда объем центрального «кубика» обращается в нуль, получается также строго симметричная структура из 12 плоских треугольников с общей вершиной в центре. Этого в самом деле можно добиться. Но возникающее предельное положение равновесия не является устойчивым: внезапно оно сменяется одним из трех положений, изображенных на рис. 240. Все явления можно наблюдать вполне отчетливо, если раствор сделать несколько более вязким, чем было указано в нашем рецепте. Перед нами возникает яркая картина, показывающая, что даже в проблемах из области физики решение не всегда находится в непрерывной зависимости от начальных данных: в самом деле, в предельном случае, когда объем воздуха, заключенного в «кубическом» пузыре, обращается в нуль, решение, изображенное на рис. 240,

не является предельным для цепи решений, изображенных на рис. 258, возникающих для различных объемов ε, когда ε стремится к нулю.

ГЛАВА VIII

Математический анализ

Введение

Было бы слишком большим упрощением представлять себе, что математический анализ «изобретен» двумя людьми: Ньютоном и Лейбницем. В действительности он сложился в итоге долгой эволюции, которая не была ни начата, ни закончена Ньютоном или Лейбницем, но в которой они оба сыграли значительную роль. Несколько математиков-энтузиастов из разных стран Европы в XVII в. поставили своей целью продолжение математической работы Галилея и Кеплера. Эти люди поддерживали друг с другом тесное общение с помощью переписки и личных встреч. Внимание их было привлечено двумя центральными проблемами. Во-первых, *проблемой касательной*: определить касательную к данной кривой — основная задача дифференциального исчисления. Во-вторых, *проблемой квадратуры*: определить площадь, связанную с заданной кривой, — основная задача интегрального исчисления. Величайшей заслугой Ньютона и Лейбница является то, что они ясно осознали *внутреннюю связь между этими двумя проблемами*. И вот объединенный таким образом метод сделался в их руках мощным орудием науки. В значительной степени успех был обусловлен поистине чудесными символическими обозначениями, придуманными Лейбницем. Заслуги этого ученого нисколько не умаляются тем, что им руководили смутные неуловимые идеи, такие идеи, которые иной раз способны заменить недостаток точного понимания в умах, предпочитающих мистицизм ясности. Ньютон, деятель точной науки в подлинном смысле слова, был, по-видимому, главным образом вдохновляем своим учителем и предшественником по Кембриджу Барроу (1630–1677), Лейбниц же пришел к математике скорее со стороны. Блестящий знаток законов, дипломат и философ, один из самых деятельных и многосторонних умов своего века, он изучил новейшую математику в невероятно короткое время у Гюйгенса, физика по специальности, во время своего пребывания в Париже в дипломатической миссии. Вскоре после этого он опубликовал результаты, которые содержат в себе ядро современного анализа. Ньютон, открытия которого были сделаны много раньше, не был расположен

их опубликовывать. Более того, хотя первоначально многие результаты, содержащиеся в его несравненном произведении «Principia», он нашел с помощью методов анализа, изложить их он предпочел в стиле классической геометрии; таким образом, в «Principia» почти совсем нет явных следов анализа. Лишь позднее были опубликованы его работы о методе «флюксий». Его почитатели вступили в жестокую схватку из-за приоритета с друзьями Лейбница. Они обвиняли последнего в плагиате, хотя трудно себе представить что-либо более естественное, чем одновременное и независимое открытие, когда атмосфера уже насыщена элементами какой-нибудь новой теории. Последовавшие пререкания по поводу «изобретения» анализа служат грустным примером того, как переоценивание вопросов о первенстве способно отравить атмосферу естественного научного единения.

Настоящая глава должна быть рассматриваема как элементарное введение, имеющее своей целью в гораздо большей степени познакомить читателя с основными концепциями, чем научить формальным операциям. Мы будем здесь широко применять «интуитивный язык», но при этом позаботимся, чтобы он не оказывался в противоречии с точными понятиями и научно обоснованными операциями.

§ 1. Интеграл

1. Площадь как предел. Для того чтобы вычислить площадь плоской фигуры, мы в качестве *единицы площади* выбираем квадрат со стороной, равной единице длины. Если единицей длины является сантиметр, соответствующей единицей площади будет квадратный сантиметр, т. е. квадрат, длина стороны которого равна сантиметру. С помощью этого определения весьма легко вычислить площадь прямоугольника. Если длины двух смежных сторон, измеренные в линейных единицах, представляются числами p и q, то площадь прямоугольника равна pq квадратных единиц или, короче, площадь равна произведению pq. Это справедливо для любых p и q, как рациональных, так и иррациональных. В случае рациональных значений p и q мы получаем этот результат, выполняя замену $p = \frac{m}{n}$, $q = \frac{m'}{n'}$, где m, m' — целые числа, а n', n — натуральные. После этого мы находим общую меру $\frac{1}{N} = \frac{1}{nn'}$ обеих сторон — таким образом, что $p = mn' \cdot \frac{1}{N}$, $q = nm' \cdot \frac{1}{N}$. Наконец, мы разбиваем прямоугольник на мелкие квадратики со стороной $\frac{1}{N}$ и с площадью $\frac{1}{N^2}$. Всего таких квадратиков будет $nm' \cdot mn'$, и общая площадь равна $nm' \cdot mn' \cdot \frac{1}{N^2} = \frac{nm' \cdot mn'}{n^2 n'^2} = \frac{m}{n} \cdot \frac{m'}{n'} = pq$. Для случая иррациональных p и q тот же результат

получится, если сначала заменим p и q соответственно приближающими их рациональными числами p_r и q_r, а затем заставим p_r и q_r стремиться к p и q.

Геометрически очевидно, что площадь треугольника равна половине площади прямоугольника с тем же основанием b и высотой h; таким образом, площадь треугольника выражается хорошо известной формулой $\frac{1}{2}\,bh$. Любая плоская область, ограниченная одной или несколькими ломаными, может быть разбита на треугольники; таким образом, ее площадь может быть получена как сумма площадей этих треугольников.

Потребность в более общем методе вычисления площадей возникает в связи с вопросом о вычислении площадей фигур, ограниченных уже не ломаными, а кривыми. Каким образом станем мы определять, например, площадь круга или сегмента параболы? Этот капитальной важности вопрос, с решением которого связано обоснование интегрального исчисления, рассматривался с очень давних пор; еще в III в. до нашей эры Архимед вычислял площади подобного рода с помощью процедуры «исчерпания». Попробуем вместе с Архимедом и великими математиками до времен Гаусса стать на «наивную» точку зрения, согласно которой криволинейные площади являются *интуитивно данными сущностями*, так что вопрос стоит не об *определении* понятия площади, а о вычислении площади (см., однако, анализ понятия, произведенный на стр. 489). В рассматриваемую криволинейную область впишем многоугольник, ограниченный ломаной линией и обладающий прекрасно определенной площадью. Выбирая новый многоугольник такого же типа, включающий первый, мы получим лучшее приближение для площади заданной области. Продолжая таким образом, мы постепенно «исчерпаем» всю область и получим искомую площадь как предел площадей надлежащим образом подобранной последовательности вписанных многоугольников с возрастающим числом сторон. Так может быть вычислена площадь круга с радиусом 1; ее числовое значение обозначается символом π.

Эту общую схему Архимед провел до конца в случае круга и в случае параболического сегмента. В течение XVII столетия было с успехом разобрано много других примеров. В каждом случае само вычисление предела ставилось в зависимость от того или иного остроумного приема, специально подобранного для каждой отдельной задачи. Одним из главных достижений анализа была замена этих специальных искусственных процедур одним общим и мощным методом.

2. Интеграл. Первым основным понятием анализа является понятие интеграла. В этой главе мы будем понимать интеграл как *площадь под кривой*, выраженную с помощью предела. Пусть дана непрерывная положительная функция $y = f(x)$, например $y = x^2$ или $y = 1 + \cos x$;

рассмотрим область, ограниченную снизу отрезком оси от некоторой точки a до некоторой точки b (причем a меньше b), справа и слева — перпендикулярами к оси x в этих точках, а сверху — кривой $y = f(x)$. Цель наша — вычислить площадь A этой области.

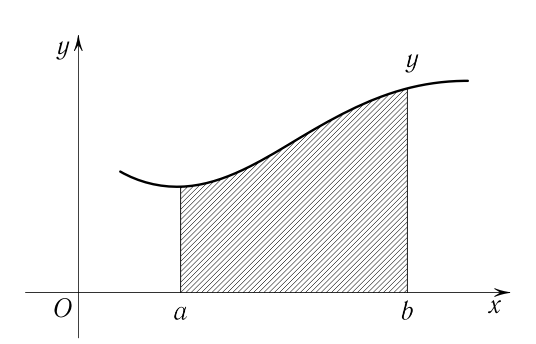


Рис. 259. Интеграл как площадь

Так как такая площадь не может быть, вообще говоря, разбита на прямоугольники или треугольники, то непосредственно нельзя указать точную математическую формулу, которая была бы пригодна для вычисления площади A. Но мы можем находить приближенные значения для A и, следовательно, представить A как предел следующим образом: разделим промежуток от $x = a$ до $x = b$ на некоторое число маленьких частных промежутков, восставим перпендикуляры в каждой точке деления, и каждую полоску области под кривой заменим прямоугольником, высоту которого выберем произвольно между наибольшей и наименьшей ординатами кривой в этой полоске. Сумма S площадей этих прямоугольников даст приближенное значение истинной площади «под» данной кривой. Точность этого приближения тем лучше, чем больше число прямоугольников и чем меньше ширина каждой отдельной полоски. Итак, мы принимаем следующее определение интересующей нас площади: *если мы построим последовательность*

$$S_1, \quad S_2, \quad S_3, \quad \ldots \tag{1}$$

приближений прямоугольниками площади под кривой, причем основание самого широкого прямоугольника в сумме S_n *стремится к* 0, *когда* n *возрастает, то последовательность* (1) *стремится к пределу* A:

$$S_n \to A, \tag{2}$$

и этот предел A, *представляющий собой площадь под данной кривой, не зависит от того, каким именно образом выбрана последовательность* (1), *раз только основания прямоугольников неограниченно уменьшаются.* (Например, S_n может произойти из S_{n-1} путем прибавления одной или нескольких новых точек к прежним, определяющим S_{n-1}, или же выбор точек деления для S_n может совершенно не зависеть от выбора точек для S_{n-1}.) Площадь A данной области, выраженную указанным предельным переходом, мы называем, по определению, *ин-*

тегралом от функции $f(x)$ в пределах от a до b. Вводя специальный символ — знак интеграла, запишем это так:

$$A = \int\limits_a^b f(x)\,dx. \tag{3}$$

Символ $\int$, значок dx и название «интеграл» были введены Лейбницем, чтобы намекнуть на способ получения этого предела. Чтобы объяснить это обозначение, мы еще раз, с бóльшими подробностями, повторим процесс приближения площади A. И в то же время аналитическая формулировка перехода к пределу позволит отбросить стесняющие предположения $f(x) \geqslant 0$ и $b > a$ и в конце концов избавиться от первоначальной интуитивной концепции интеграла как «площади под кривой» (это будет сделано в дополнении, § 1).

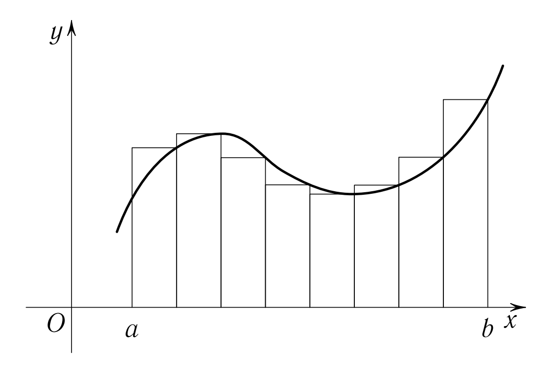


Рис. 260. Приближение площади ступенчатой фигурой

Разделим промежуток от a до b на n маленьких частных промежутков, которые только ради простоты мы будем предполагать имеющими одинаковую длину $\dfrac{b-a}{n}$; обозначим точки деления следующим образом:

$$x_0 = a, \quad x_1 = a + \frac{b-a}{n}, \quad x_2 = a + \frac{2(b-a)}{n}, \ \ldots, \ x_n = a + \frac{n(b-a)}{n} = b.$$

Введем для обозначения величины $\dfrac{b-a}{n}$ разности между двумя последовательными значениями x символ Δx (читается «дельта икс»):

$$\Delta x = \frac{b-a}{n} = x_{j+1} - x_j,$$

где символ Δ обозначает просто «разность». Это оперативный символ, который нельзя рассматривать как числовой множитель. За высоту каждого приближающего прямоугольника мы можем принять значение $y =$

$f(x)$ в правой крайней точке соответствующего промежутка. Тогда сумма площадей прямоугольников будет равна

$$S_n = f(x_1)\Delta x + f(x_2)\Delta x + \ldots + f(x_n)\Delta x, \tag{4}$$

или, сокращенно,

$$S_n = \sum_{j=1}^{n} f(x_j)\Delta x. \tag{5}$$

Символ $\sum\limits_{j=1}^{n}$ (читается «сумма по j от 1 до n») обозначает сумму всех выражений, получаемых, когда j последовательно пробегает значения 1, 2, 3, ..., n.

Употребление символа $\sum$ для выражения результата суммирования в сжатой форме можно иллюстрировать следующими примерами:

$$2 + 3 + 4 + \ldots + 10 = \sum_{j=2}^{10} j,$$

$$1 + 2 + 3 + \ldots + n = \sum_{j=1}^{n} j,$$

$$1^2 + 2^2 + 3^2 + \ldots + n^2 = \sum_{j=1}^{n} j^2,$$

$$aq + aq^2 + \ldots + aq^n = \sum_{j=1}^{n} aq^j,$$

$$a + (a + d) + (a + 2d) + \ldots + (a + nd) = \sum_{j=0}^{n} (a + jd).$$

Построим теперь последовательность таких приближений S_n, в которых n возрастает неограниченно, так что число членов в каждой из сумм (5) стремится к бесконечности, в то время как каждый отдельный член $f(x_j)\Delta x$ стремится к 0 вследствие присутствия множителя $\Delta x = \dfrac{b-a}{n}$. При возрастании n эта сумма стремится к площади A:

$$A = \lim_{n\to\infty} \sum_{j=1}^{n} f(x_j)\Delta x = \int\limits_a^b f(x)\,dx. \tag{6}$$

Лейбниц символизировал этот предельный переход от приближающих сумм S_n к пределу A заменой знака суммирования $\sum$ через $\int$, а символа разности Δ символом d. (Во времена Лейбница знак суммирования $\sum$ писался обычно в виде S, и символ $\int$ представляет собой просто стилизацию буквы S.) Несмотря на то что символика Лейбница хорошо намекает на способ, каким был получен интеграл, не следует придавать

слишком большого значения тому, что является лишь чисто условным приемом обозначения предела. В ранние дни анализа, когда отчетливого понятия о пределах еще не существовало и необходимость предельного перехода часто упускалась из виду, многие пытались объяснить смысл интеграла, говоря, что «конечное приращение Δx заменено бесконечно малой величиной dx, а сам интеграл есть сумма бесконечно большого числа бесконечно малых слагаемых $f(x)\,dx$». Хотя «бесконечно малое» имеет известную притягательность для умов, имеющих склонность к философским спекуляциям, ему нет и не может быть места в современной математике. Никакой полезной цели нельзя достигнуть, окружая ясное понятие интеграла туманом не имеющих смысла фраз. Но даже сам Лейбниц иногда поддавался соблазнительному воздействию своих символов: в самом деле, они работают так, *как если бы* обозначали сумму «бесконечно малых» величин, с которыми можно до некоторой степени оперировать, как с обыкновенными величинами. Даже само слово «интеграл» было создано для того, чтобы обозначить, что «всё», т. е. «полная» площадь A, составлено из «бесконечно малых» частиц $f(x)\,dx$. Как бы то ни было, прошло около сотни лет после Ньютона и Лейбница, прежде чем было ясно осознано, что истинной основой определения интеграла является понятие предела, и ничего больше. Твердо став на эту точку зрения, мы избежим всякой неясности, всех трудностей и всех нелепостей, которые вызывали такое смущение в ранний период развития анализа.

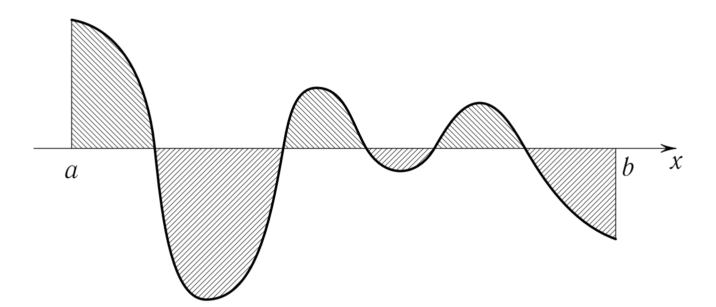


Рис. 261. Положительные и отрицательные площади

3. Общие замечания о понятии интеграла. Общее определение. В нашем геометрическом определении интеграла как площади мы явно предполагали, что функция $f(x)$ во всем промежутке интегрирования $[a,\,b]$ неотрицательна, т. е. что никакая часть графика не лежит под осью x. В аналитическом же определении интеграла как предела последовательности сумм S_n такое предположение является излишним. Мы просто возьмем малые количества $f(x_j)\Delta x$, составим их сумму и перейдем к пределу; эта процедура остается имеющей вполне опреде-

ленный смысл и в том случае, если некоторые или все значения $f(x_j)$ отрицательны. Интерпретируя это геометрически с помощью площадей (рис. 261), мы приходим к заключению, что интеграл от $f(x)$ представляет собой алгебраическую сумму площадей, ограниченных графиком и осью x, причем площади, лежащие под осью x, считаются отрицательными, остальные — положительными.

Может случиться, что в тех или иных случаях мы придем к интегралам $\int\limits_a^b f(x)\,dx$, в которых b меньше, чем a, так что $\dfrac{b-a}{n} = \Delta x$ окажется отрицательным числом. Тогда в нашем аналитическом определении члены вида $f(x_j)\Delta x$ будут отрицательными, если $f(x_j)$ положительно, а Δx отрицательно, и т. д. Другими словами, величина этого интеграла будет только знаком отличаться от величины интеграла в пределах от b до a. Таким образом получаем следующее простое свойство интеграла:

$$\int\limits_a^b f(x)\,dx = -\int\limits_b^a f(x)\,dx.$$

Далее, нужно подчеркнуть, что значение интеграла не меняется и в том случае, если точки деления x_j не будут выбираться равноотстоящими, другими словами, если разности $\Delta x = x_{j+1} - x_j$ не будут одинаковы. Мы можем выбрать x_j произвольно, и тогда разности $\Delta x = x_{j+1} - x_j$ должны быть различаемы с помощью соответствующих значков. Даже в этом предположении сумма

$$S_n = f(x_1)\Delta x_0 + f(x_2)\Delta x_1 + \ldots + f(x_n)\Delta x_{n-1},$$

а также сумма

$$S_n' = f(x_0)\Delta x_0 + f(x_1)\Delta x_1 + \ldots + f(x_{n-1})\Delta x_{n-1}$$

будут стремиться к одному и тому же пределу, именно к значению интеграла $\int\limits_a^b f(x)\,dx$, если только мы позаботимся о том, чтобы с возрастанием x все разности $\Delta x_j = x_{j+1} - x_j$ стремились к нулю таким образом, чтобы наибольшая из них (при данном значении n) стремилась к нулю, когда n неограниченно возрастает.

Окончательное определение интеграла дается с помощью формулы

$$\int\limits_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{j=1}^{n} f(v_j)\Delta x_j. \tag{6a}$$

Под знаком суммы число v_i может обозначать *любую* точку в промежутке $x_j \leqslant v_j \leqslant x_{j+1}$, и единственное ограничение, касающееся способа разбиений основного промежутка, заключается в том, чтобы наибольшая из разностей $\Delta x_j = x_{j+1} - x_j$ стремилась к нулю, когда n стремится к бесконечности.

Существование предела (6а) не требует доказательства, если мы допустим как само собой разумеющееся понятие «площади под кривой», а также и возможность приближения этой площади с помощью прямоугольников. И все же, как это выяснится из дальнейших рассуждений (стр. 489), более глубокий анализ показывает, что для того, чтобы определение интеграла было логически совершенным, желательно и даже необходимо доказать существование этого предела независимо от первоначального геометрического представления о площади и притом какова бы ни была непрерывная функция $f(x)$.

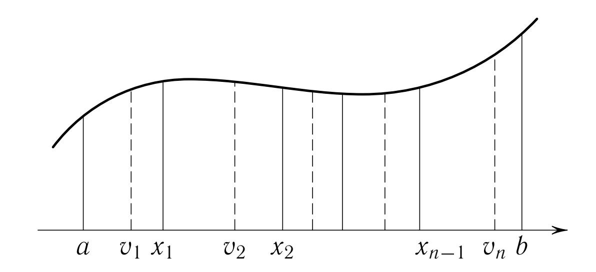


Рис. 262. Произвольность разбиения области определения функции при общем определении интеграла

4. Примеры интегрирования. Интегрирование функции x^r. До сих пор наши рассуждения об интеграле были чисто теоретическими. Возникает основной вопрос по поводу рассмотренного построения сумм S_n по общей установленной схеме и последующего перехода к пределу: ведет ли эта процедура к каким-либо осязаемым результатам в отдельных конкретных случаях? Конечно, решение этого вопроса потребует некоторых дополнительных рассуждений, приспособленных к тем специальным функциям $f(x)$, от которых нужно найти интеграл.

Когда Архимед две тысячи лет назад вычислил площадь параболического сегмента, он выполнил то, что мы теперь называем интегрированием функции $f(x) = x^2$, притом чрезвычайно остроумным способом; в XVII столетии предшественники Ньютона и Лейбница успешно решили проблему интегрирования таких простых функций, как x^n, опять-таки с помощью специальных приемов. Только после рассмотрения большого числа конкретных примеров был найден общий подход к проблеме интегрирования на основе систематического метода, и таким образом область разрешимых задач была сильно расширена. В настоящей главе мы рассмотрим небольшое число отдельных конструктивных задач, принадлежащих к эпохе «праанализа», так как для операции интегрирования, понимаемой как предельный процесс, лучшей иллюстрации не

придумаешь.

а) Начнем с совершенно тривиального примера. Если $y = f(x)$ является константой, например, $f(x) = 2$, то, очевидно, интеграл $\int\limits_a^b 2\,dx$, понимаемый как площадь, равен $2(b-a)$, поскольку площадь прямоугольника равна произведению основания на высоту. Сравним этот результат с определенным интегралом. Если в формуле (5) мы подставим $f(x_j) = 2$ для всех значений j, то при любом значении n найдем, что

$$S_n = \sum_{j=1}^{n} f(x_j)\Delta x = \sum_{j=1}^{n} 2\Delta x = 2\sum_{j=1}^{n} \Delta x = 2(b-a);$$

в самом деле,

$$\sum_{j=1}^{n} \Delta x = (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_n - x_{n-1}) = x_n - x_0 = b - a.$$

б) Почти так же просто проинтегрировать функцию $f(x) = x$. В этом примере интеграл $\int\limits_a^b x\,dx$ является площадью трапеции (рис. 263), следовательно, согласно элементарной геометрии выразится формулой

$$(b-a)\,\frac{b+a}{2} = \frac{b^2 - a^2}{2}.$$

Этот же результат получается и из определения интеграла (6), в чем можно убедиться фактическим переходом к пределу без обращения к геометрическому представлению: если мы в формуле (5) положим $f(x) = x$, то сумма S_n примет вид

$$\begin{aligned} S_n = \sum_{j=1}^{n} x_j\,\Delta x = \sum_{j=1}^{n} (a + j\Delta x)\Delta x = \\ = (na + \Delta x + 2\Delta x + 3\Delta x + \ldots + n\Delta x)\Delta x = \\ = na\Delta x + (\Delta x)^2(1 + 2 + 3 + \ldots + n). \end{aligned}$$

Применяя формулу для суммы арифметической прогрессии $1 + 2 + 3 + \ldots + n$, выведенную на стр. 31, формула (1), мы получим

$$S_n = na\Delta x + \frac{n(n+1)}{2}(\Delta x)^2.$$

И так как

$$\Delta x = \frac{b-a}{n},$$

то отсюда следует

$$S_n = a(b-a) + \frac{1}{2}(b-a)^2 + \frac{1}{2n}(b-a)^2.$$

Пусть теперь n стремится к бесконечности; тогда переход к пределу даст результат

$$\lim_{n\to\infty} S_n = \int\limits_a^b x\,dx = a(b-a) + \frac{1}{2}(b-a)^2 = \frac{1}{2}(b^2 - a^2),$$

в полном соответствии с геометрической интерпретацией интеграла как площади.

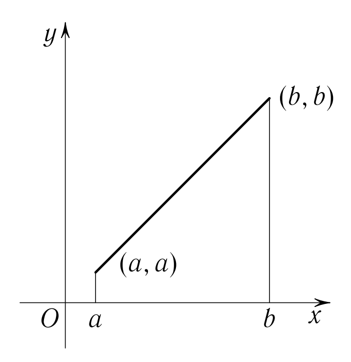


Рис. 263. Площадь трапеции

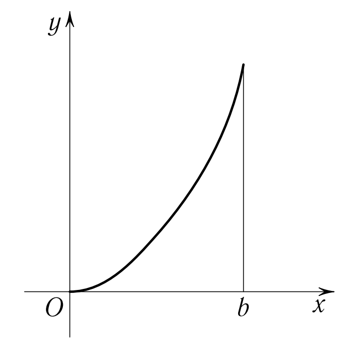


Рис. 264. Площадь под параболой

в) Менее тривиальным является интегрирование функции $f(x) = x^2$. Архимед употребил геометрический метод при решении эквивалентной задачи — нахождении площади сегмента параболы $y = x^2$. Здесь мы будем действовать аналитически, базируясь на определении (6а). Чтобы упростить формальные выкладки, в качестве «нижнего предела» интеграла a выберем 0; тогда $\Delta x = \dfrac{b}{n}$. Так как $x_j = j \cdot \Delta x$ и $f(x_j) = j^2(\Delta x)^2$, то для суммы S_n мы получим выражение

$$S_n = \sum_{j=1}^{n} (j\Delta x)^2 \Delta x = [1^2 \cdot (\Delta x)^2 + 2^2 \cdot (\Delta x)^2 + \ldots + n^2 \cdot (\Delta x)^2] \cdot \Delta x =$$
$$= (1^2 + 2^2 + \ldots + n^2) \cdot (\Delta x)^3.$$

Теперь можно фактически вычислить предел. Применяя формулу

$$1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6},$$

установленную на стр. 33, и заменяя Δx через $\dfrac{b}{n}$, мы получим

$$S_n = \frac{n(n+1)(2n+1)}{6} \cdot \frac{b^3}{n^3} = \frac{b^3}{6}\left(1 + \frac{1}{n}\right)\left(2 + \frac{1}{n}\right).$$

Это предварительное преобразование облегчает предельный переход: при неограниченном возрастании n обратная величина $\dfrac{1}{n}$ стремится к

нулю, и потому в качестве предела получается просто $\frac{b^3}{6} \cdot 1 \cdot 2 = \frac{b^3}{3}$; следовательно, окончательный результат имеет вид

$$\int_0^b x^2\,dx = \frac{b^3}{3}.$$

Применяя этот результат к площади от 0 до a, получим

$$\int_0^a x^2\,dx = \frac{a^3}{3};$$

наконец, вычитание площадей дает

$$\int_a^b x^2\,dx = \frac{b^3 - a^3}{3}.$$

Упражнение. Тем же способом, употребляя формулу (5) со стр. 33, докажите, что

$$\int_a^b x^3\,dx = \frac{b^4 - a^4}{4}.$$

Применяя общие формулы для сумм $1^k + 2^k + \ldots + n^k$ k-х степеней целых чисел от 1 до n, можно было бы получить результат

$$\int_a^b x^k\,dx = \frac{b^{k+1} - a^{k+1}}{k+1} \tag{7}$$

при любом целом положительном значении k.

* Вместо того чтобы действовать этим путем, мы можем получить несколько проще даже более общий результат, воспользовавшись сделанным раньше замечанием о возможности вычислить интеграл и при неравноотстоящих точках деления. Мы выведем формулу (7) не только для любого целого положительного k, но и для любого положительного или отрицательного рационального числа

$$k = \frac{u}{v},$$

где u — целое положительное, а v — целое положительное или отрицательное число. Исключается только значение $k = -1$, при котором формула (7) теряет смысл. Предположим также, что $0 < a < b$.

Чтобы получить формулу (7), построим сумму S_n, выбирая точки деления $x_0 = a$, x_1, x_2, ..., $x_n = b$ в геометрической прогрессии. Положим $\sqrt[n]{\frac{b}{a}} = q$, так что $\frac{b^n}{a^n} = q^n$, и определим: $x_0 = a$, $x_1 = aq$, $x_2 = aq^2$, ..., $x_n = aq^n = b$. При таком выборе значений x_j, как мы увидим, предельный переход совершается особенно просто. Поскольку $f(x_j) = x_j^k = a^k q^{jk}$ и $\Delta x_j = x_{j+1} - x_j = aq^{j+1} - aq^j$, мы будем иметь выражение

$$S_n = a^k(aq - a) + a^k q^k(aq^2 - aq) + a^k q^{2k}(aq^3 - aq^2) + a^k q^{(n-1)k}(aq^n - aq^{n-1}).$$

Так как каждый член содержит множители $a^k(aq-a)$, то можно написать

$$S_n = a^{k+1}(q-1)\{1+q^{k+1}+q^{2(k+1)}+\ldots+q^{(n-1)(k+1)}\}.$$

Подставляя t вместо q^{k+1}, видим, что выражение в скобках является геометрической прогрессией $1+t+t^2+\ldots+t^{n-1}$, сумма которой, как показано на стр. 32, равна $\frac{t^n-1}{t-1}$. Но

$$t^n = q^{n(k+1)} = \left(\frac{b}{a}\right)^{k+1} = \frac{b^{k+1}}{a^{k+1}}.$$

Таким образом,

$$S_n = (q-1)\frac{b^{k+1}-a^{k+1}}{q^{k+1}-1} = \frac{b^{k+1}-a^{k+1}}{N}, \tag{8}$$

где

$$N = \frac{q^{k+1}-1}{q-1}.$$

До сих пор n было фиксированным числом. Пусть теперь n возрастает; определим тогда предел, к которому стремится N. При возрастании n корень $\sqrt[n]{\frac{b}{a}} = q$ стремится к 1 (см. стр. 346); поэтому и числитель и знаменатель выражения N стремятся к нулю, что побуждает к осторожности. Предположим сначала, что k — целое положительное число, тогда можно осуществить деление на $q-1$, и мы получим (см. стр. 120): $N = q^k + q^{k-1} + \ldots + q + 1$. Если теперь n возрастает, так что q стремится к 1, а следовательно, и q^2, q^3, ..., q^k также стремятся к 1, то N стремится к $k+1$. Но из этого вытекает, что S_n стремится к $\frac{b^{k+1}-a^{k+1}}{k+1}$, что и требовалось доказать.

Упражнение. Докажите, что при любом рациональном $k \neq -1$ остается в силе та же самая предельная формула $N \to k+1$, а следовательно, сохраняется и результат (7). Сначала дайте доказательство, следуя нашему образцу, в предположении, что k целое отрицательное. Затем, если $k = \frac{u}{v}$, положите $q^{\frac{1}{v}} = s$, откуда следует

$$N = \frac{s^{(k+1)v}-1}{s^v-1} = \frac{s^{u+v}-1}{s^v-1} = \frac{s^{u+v}-1}{s-1} : \frac{s^v-1}{s-1}.$$

Если n возрастает, так что q и s стремятся к 1, то отношения в последней части равенства стремятся соответственно к $u+v$ и к v, что в качестве предела для N снова дает $\frac{u+v}{v} = k+1$.

В § 5 мы увидим, каким образом с помощью мощных методов анализа можно упростить это длинное и несколько искусственное рассуждение.

Упражнения. 1) Проверьте предшествующее интегрирование x^k для случаев $k = \frac{1}{2}, -\frac{1}{2}, 2, -2, 3, -3$.

2) Вычислите значения интегралов:

а) $\int\limits_{-2}^{-1} x\,dx$, б) $\int\limits_{-1}^{+1} x\,dx$, в) $\int\limits_{1}^{2} x^2\,dx$, г) $\int\limits_{1}^{-2} x^3\,dx$, д) $\int\limits_{0}^{n} x\,dx$.

3) Найдите значения интегралов:

а) $\int\limits_{-1}^{+1} x^3\,dx$, б) $\int\limits_{-2}^{2} x^3\cos x\,dx$, в) $\int\limits_{-1}^{+1} x^4\cos^2 x\sin^5 x\,dx$, г) $\int\limits_{-1}^{+1} \operatorname{tg} x\,dx$.

(*Указание*: рассмотрите графики функций под знаком интеграла, принимая во внимание их симметрию по отношению к оси $x = 0$, и интерпретируйте интегралы как площади.)

*4) Проинтегрируйте $\sin x$ и $\cos x$ в пределах от 0 до b, подставляя h вместо Δx и применяя формулы со стр. 512.

5) Проинтегрируйте $f(x) = x$ и $f(x) = x^2$ в пределах от 0 до b, производя раздробление на равные части и в формуле (6а) выбирая значения $v_j = \dfrac{x_j + x_{j+1}}{2}$.

*6) Пользуясь результатом (7) и определением интеграла с равными значениями для Δx, докажите предельное соотношение

$$\frac{1^k + 2^k + \ldots + n^k}{n^{k+1}} \to \frac{1}{k+1} \quad \text{при} \quad n \to \infty.$$

$\Big($*Указание*: положите $\dfrac{1}{n} = \Delta x$ и покажите, что рассматриваемый предел равен интегралу $\int\limits_0^1 x^k\,dx.\Big)$

*7) Докажите, что при $n \to \infty$ справедливо следующее предельное соотношение:

$$\frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{1+n}} + \frac{1}{\sqrt{2+n}} + \ldots + \frac{1}{\sqrt{n+n}}\right) \to 2(\sqrt{2} - 1).$$

(*Указание*: напишите эту сумму так, чтобы ее предел являлся некоторым интегралом.)

8) Вычислите площадь параболического сегмента, ограниченного дугой P_1P_2 и хордой P_1P_2 параболы $y = ax^2$, выражая результат через координаты точек P_1 и P_2.

5. Правила «интегрального исчисления». Важной ступенью в развитии интегрального исчисления явилось формулирование некоторых общих правил, с помощью которых более сложные задачи могут сводиться к более простым, а тем самым могут быть решены почти механически. Алгоритмический характер этих правил особенно ярко подчеркивается обозначениями Лейбница. Однако слишком сосредоточивать внимание на механизме решения задач при изучении анализа значило бы снижать содержание предмета и могло бы привести к пустой долбежке.

Некоторые простые правила интегрирования следуют сразу или из определения (6), или из геометрической интерпретации интегралов как площадей.

Интеграл от суммы двух функций равен сумме интегралов от этих двух функций. Интеграл от произведения функции на постоянное c равен произведению интеграла от функции на постоянное c. Эти два правила можно выразить одной формулой

$$\int_a^b [cf(x) + kg(x)]\,dx = c\int_a^b f(x)\,dx + k\int_a^b g(x)\,dx. \tag{9}$$

Доказательство следует непосредственно из определения интеграла как предела конечных сумм (5), поскольку соответствующая формула для суммы S_n, очевидно, справедлива. Это правило обобщается тотчас же на сумму более чем двух функций.

В качестве примера применения этого правила рассмотрим полином

$$f(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n,$$

коэффициенты которого a_0, a_1, ..., a_n постоянны. Чтобы вычислить интеграл от функции $f(x)$ в пределах от a до b, мы будем интегрировать почленно, согласно правилу. Применяя формулу (7), мы найдем

$$\int_a^b f(x)\,dx = a_0(b-a) + a_1\,\frac{b^2-a^2}{2} + \ldots + a_n\,\frac{b^{n+1}-a^{n+1}}{n+1}.$$

Другое правило, вытекающее со всей очевидностью как из аналитического определения интеграла, так и из его геометрической интерпретации, выражается формулой

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx. \tag{10}$$

Кроме того, ясно, что наш основной интеграл равен нулю, если b равно a. Правило

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx, \tag{11}$$

приведенное на стр. 428, не стоит в противоречии с последними двумя, поскольку оно получается из (10) при $c = a$.

Иногда бывает удобно использовать то обстоятельство, что значение интеграла не зависит от выбора наименования независимого переменного интегрируемой функции; например,

$$\int_a^b f(x)\,dx = \int_a^b f(u)\,du = \int_a^b f(t)\,dt \quad \text{и т. д.}$$

В самом деле, простая замена наименований координат, к системе которых отнесен график функции, не меняет площади под данной кривой.

Аналогичное замечание относится к тому случаю, когда производится некоторая замена в самой системе координат. Например, перенесем начало координат на одну единицу вправо из точки O в точку O', как показано на рис. 265, таким образом, что x будет заменено новой координатой x' по формуле $x = 1 + x'$. Уравнение кривой $y = f(x)$ в новой системе координат примет вид $y = f(1 + x')$ $\left(\text{например, } y = \dfrac{1}{x} = \dfrac{1}{1+x'}\right)$. Данная площадь A под этой кривой, скажем, в пределах от $x = 1$ до $x = b$, в новой системе координат будет площадью под кривой в пределах от $x' = 0$ до $x' = b - 1$. Таким образом, будем иметь

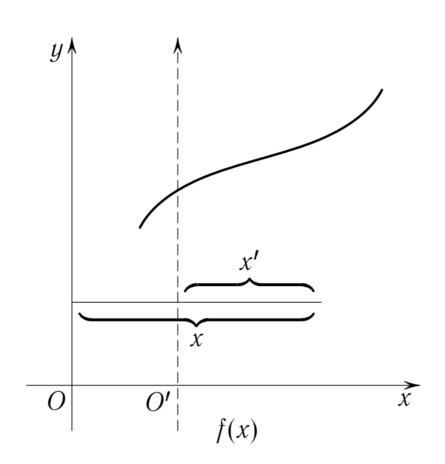


Рис. 265. Перемещение оси y

$$\int_1^b f(x)\,dx = \int_0^{b-1} f(1+x')\,dx'$$

и, написав букву u вместо буквы x', получим

$$\int_1^b f(x)\,dx = \int_0^{b-1} f(1+u)\,du; \tag{12}$$

например,

$$\int_1^b \frac{1}{x}\,dx = \int_0^{b-1} \frac{1}{1+u}\,du; \tag{12а}$$

а для функции $f(x) = x^k$ получим таким же образом

$$\int_1^b x^k\,dx = \int_0^{b-1} (1+u)^k\,du. \tag{12б}$$

Аналогично,

$$\int_0^b x^k\,dx = \int_{-1}^{b-1} (1+u)^k\,du \quad (k \geqslant 0), \tag{12в}$$

и, поскольку интеграл в левой части (12в) равен $\dfrac{b^{k+1}}{k+1}$, мы получим

$$\int_{-1}^{b-1} (1+u)^k\,du = \frac{b^{k+1}}{k+1}. \tag{12г}$$

Упражнения. 1) Вычислите интеграл от многочлена $1 + x + x^2 + \ldots + x^n$ в пределах от 0 до b.

2) Докажите при $n > 0$, что интеграл от функции $(1+x)^n$ в пределах от -1 до z равен дроби

$$\frac{(1+z)^{n+1}}{n+1}.$$

3) Покажите, что интеграл от $x^n \sin x$ в пределах от 0 до 1 меньше, чем $\frac{1}{n+1}$. (*Указание*: последняя величина есть значение интеграла от x^n.)

4) Докажите непосредственно, а также пользуясь разложением по формуле бинома, что интеграл от функции $\frac{(1+x)^n}{n}$ в пределах от -1 до z равен дроби $\frac{(1+z)^{n+1}}{n(n+1)}$.

Следует упомянуть, наконец, два важных правила, которые выражаются посредством неравенств. Эти правила дают, правда, грубые, но все же полезные оценки для значения интегралов.

Предположим, что $b > a$ и что значения функции $f(x)$ в промежутке от a до b нигде не превосходят значений другой функции $g(x)$. Тогда мы имеем

$$\int_a^b f(x)\,dx \leqslant \int_a^b g(x)\,dx, \qquad (13)$$

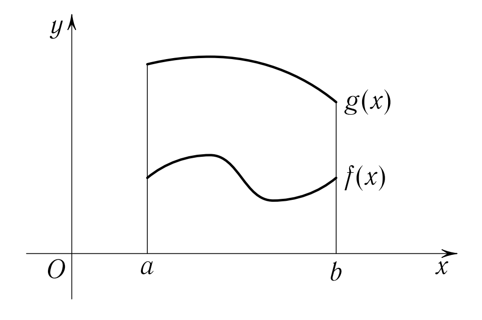


Рис. 266. Сравнение интегралов

что непосредственно ясно или из рис. 266, или из аналитического определения интеграла. В частности, если функция $g(x)$ равна M, т. е. является постоянной, то мы получаем:

$$\int_a^b g(x)\,dx = \int_a^b M\,dx = M(b-a);$$

отсюда следует неравенство

$$\int_a^b f(x)\,dx \leqslant M(b-a). \qquad (14)$$

Если функция $f(x)$ неотрицательна, то $f(x) = |f(x)|$. Если $f(x) < 0$, то $|f(x)| > f(x)$. Отсюда, полагая в неравенстве (13) $g(x) = |f(x)|$, мы получим полезную формулу:

$$\int_a^b f(x)\,dx \leqslant \int_a^b |f(x)|\,dx. \qquad (15)$$

Но поскольку $|-f(x)| = |f(x)|$, мы имеем также формулу

$$-\int\limits_a^b f(x)\,dx \leqslant \int\limits_a^b |f(x)|\,dx,$$

что дает вместе с формулой (15) более сильное неравенство, а именно

$$\left|\int\limits_a^b f(x)\,dx\right| \leqslant \int\limits_a^b |f(x)|\,dx. \tag{16}$$

§ 2. Производная

1. Производная как наклон. В то время как понятие интеграла своими корнями уходит в античную древность, другое основное понятие анализа — производная — было сформулировано только в XVII столетии знаменитым Ферма и другими. Сделанное Ньютоном и Лейбницем открытие органической связи между этими понятиями, казалось бы столь различными, способствовало небывалому развитию математической науки.

Ферма интересовался вопросом об определении наибольших и наименьших значений функции $y = f(x)$. При изучении графика функции принято называть максимумом точку, расположенную выше всех других, а минимумом — точку, расположенную ниже всех других точек в ее окрестности. На рис. 191 на стр. 365 точка B является максимумом, точка C — минимумом. Естественно при нахождении максимума или минимума использовать понятие касательной к кривой. Предположим, что график кривой нигде не образует острых углов и не обладает другими особенностями и что в каждой точке он имеет определенное направление, определяемое касательной прямой. В точках максимума или минимума касательная к кривой $y = f(x)$ должна быть параллельна оси x; в противном случае кривая около этих точек или поднималась бы, или опускалась бы. Это замечание побуждает нас заняться общим вопросом об определении направления касательной к кривой $y = f(x)$ в любой точке P этой кривой.

Чтобы охарактеризовать направление прямой в плоскости x, y, обыкновенно задается ее *наклон*, который представляет собой тангенс угла α между положительным направлением оси x и рассматриваемой прямой. Если P есть некоторая точка прямой L, продвигаемся вправо от нее до некоторой точки R, а затем вверх или вниз до точки Q, лежащей на прямой, тогда наклон L равен $\operatorname{tg}\alpha$, т. е. $\dfrac{RQ}{PR}$. Отрезок PR предполагается положительным, тогда как RQ — положительным или отрицательным в зависимости от того, будет ли он направлен вверх или вниз; таким образом, наклон дает нам подъем или падение на единицу длины по

горизонтали (при перемещении по прямой слева направо). На рис. 267 наклон первой прямой равен $\frac{2}{3}$, в то время как наклон второй прямой равен -1.

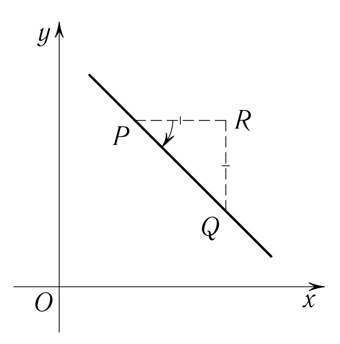


Рис. 267. Наклоны прямых

Под наклоном *кривой* в точке P мы подразумеваем наклон ее *касательной* в этой точке. Поскольку мы расположены принять понятие касательной как интуитивно данное, перед нами остается только задача — *найти способ для вычисления наклона кривой.* В настоящий момент мы встанем на именно такую точку зрения: более тщательный анализ относящихся сюда проблем будет произведен в дополнении к этой главе.

2. Производная как предел. Рассмотрение кривой $y = f(x)$ только в одной ее точке $P(x, y)$ не позволяет вычислить наклон кривой в этой точке. Необходимо прибегнуть к предельному процессу, сходному с процессом вычисления площади. Этот предельный процесс является основой дифференциального исчисления. Рассмотрим на данной кривой другую точку P_1, близкую к P, с координатами x_1, y_1; обозначим прямую, проходящую через точки P и P_1, буквой t_1; эта прямая по отношению к нашей кривой является секущей, которая мало отличается от касательной к точке P, если только точка P_1 близка к точке P. Обозначим угол между осью x и прямой t_1 буквой α_1. Заставим теперь x_1 стремиться к x; тогда точка P_1 будет двигаться по кривой к точке P и секущая t_1 будет приближаться к некоторому предельному положению, которое и есть не что иное, как касательная t к нашей кривой в точке x. Если буквой α обозначить угол между осью x и касательной t, то при $x_1 \to x$ будем иметь

$$y_1 \to y, \quad P_1 \to P, \quad t_1 \to t \quad \text{и} \quad \alpha_1 \to \alpha.$$

Касательная есть предел секущей, а наклон касательной есть предел наклона секущей[1].

[1] Наши обозначения здесь слегка отличаются от обозначений главы VI, поскольку

Хотя мы и не имеем явного выражения для наклона самой касательной t, зато наклон секущей t_1 дается формулой

$$\text{наклон } t_1 = \frac{y_1 - y}{x_1 - x} = \frac{f(x_1) - f(x)}{x_1 - x};$$

обозначая, как раньше, операцию образования разности символом Δ, мы получим

$$\text{наклон } t_1 = \frac{\Delta y}{\Delta x} = \frac{\Delta f(x)}{\Delta x}.$$

Наклон секущей t_1 есть «разностное отношение» — разность Δy значений функции, деленная на разность Δx значений независимого переменного. Сверх того, имеем наклон $t =$ предел наклона $t_1 = \lim \frac{f(x_1) - f(x)}{x_1 - x} = \lim \frac{\Delta y}{\Delta x}$, где пределы вычисляются при $x_1 \to x$, т. е. при $\Delta x = x_1 - x \to 0$.

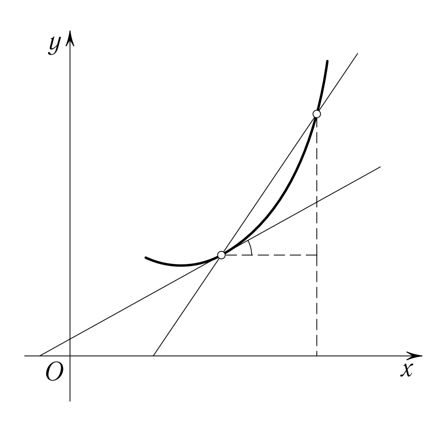


Рис. 268. Производная как предел

Касательная t к данной кривой имеет наклон, равный пределу разностного отношения $\frac{\Delta y}{\Delta x}$ при стремлении $\Delta x = x_1 - x$ к нулю.

Первоначальная функция $f(x)$ давала значение «высоты» различных точек кривой $y = f(x)$. Предположим теперь, что точка P движется по кривой $y = f(x)$. Тогда рассматриваемый наклон в точке P будет представлять некоторую новую функцию от x, которую мы обозначим через $f'(x)$ и назовем *производной* от функции $f(x)$.

Предельный процесс, с помощью которого получена производная, называется *дифференцированием* функции $f(x)$. Этот процесс есть такая операция, которая по определенному правилу сопоставляет данной функции $f(x)$ некоторую другую функцию $f'(x)$. Подобным же образом при определении самой функции $f(x)$ было установлено правило, которое сопоставляло каждому значению переменного x некоторое значение функции $f(x)$.
Итак,

$f(x)$ есть высота кривой $y = f(x)$ в точке x,

$f'(x)$ есть наклон кривой $y = f(x)$ в точке x.

Слово «дифференцирование» объясняется тем обстоятельством, что $f'(x)$ есть предел разности (differentia) $f(x_1) - f(x)$, деленной на раз-

там мы имели $x \to x_1$, где x_1 постоянно. Никакой путаницы от этого изменения обозначений не произойдет.

ность $x_1 - x$:

$$f'(x) = \lim \frac{f(x_1) - f(x)}{x_1 - x} \quad \text{при} \quad x_1 \to x. \tag{1}$$

Другим часто употребляемым обозначением является

$$f'(x) = Df(x),$$

где символ D есть первая буква слова differentia, что значит «разность»; кроме того, для производной от функции $y = f(x)$ существуют еще обозначения Лейбница

$$\frac{dy}{dx}, \quad \text{или} \quad \frac{df(x)}{dx},$$

которые мы подвергнем обсуждению в § 4, и намекающие на то, что производная получается как предел разностного отношения $\frac{\Delta y}{\Delta x}$ или $\frac{\Delta f(x)}{\Delta x}$[1].

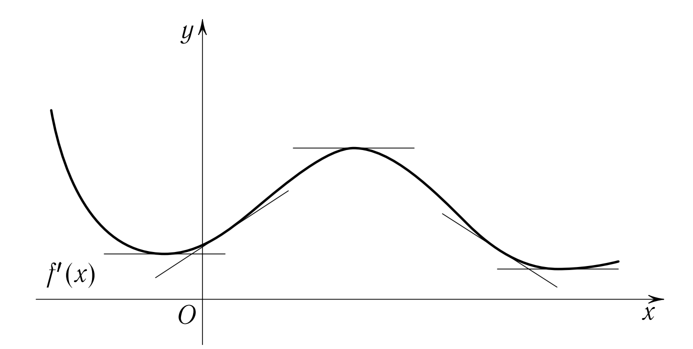


Рис. 269. Знак производной

Условившись в том, что движение по кривой совершается в направлении возрастающих значений x, мы можем теперь заключить: то обстоятельство, что производная в некоторой точке положительна, $f'(x) > 0$, обозначает подъем кривой (значения y возрастают). Напротив, то обстоятельство, что производная отрицательна, $f'(x) < 0$, означает падение кривой (значения y убывают); наконец, если производная обращается в нуль, $f'(x) = 0$, то это обозначает горизонтальное направление кривой для соответствующего значения x. В точках максимума и минимума наклон должен быть равен нулю (рис. 269). Таким образом, решая уравнение

$$f'(x) = 0$$

относительно x, мы можем найти положение максимумов и минимумов, как это и было впервые сделано Ферма.

[1] В русском издании в дальнейшем используются именно эти, общепринятые у нас обозначения. — *Прим. ред.*

3. Примеры. Может показаться, что рассуждения, приведшие к определению (1), лишены всякого практического смысла. В самом деле, одна проблема подменена другой: вместо того чтобы искать наклон касательной к кривой $y = f(x)$ в некоторой точке, мы должны вычислять предел (1), что с первого взгляда кажется одинаково трудным. Но как только мы откажемся от рассмотрении в общем виде и перейдем к отдельным функциям, мы получим весьма реальные результаты.

Простейшей из функций является функция $f(x) = c$, где c постоянно. График этой функции $y = f(x) = c$ есть горизонтальная прямая, совпадающая со всеми своими касательными; очевидно, для всех значений x имеет место соотношение

$$f'(x) = 0.$$

Это вытекает также и из определения (1); в самом деле,

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x)}{x_1 - x} = \frac{c - c}{x_1 - x} = \frac{0}{x_1 - x} = 0.$$

Таким образом, получаем тривиальный результат:

$$\lim \frac{f(x_1) - f(x)}{x_1 - x} = 0 \quad \text{при} \quad x_1 \to x.$$

Вслед за этим рассмотрим простую функцию $y = f(x) = x$, графиком которой является биссектриса угла первого квадранта. Геометрически ясно, что для всех значений x

$$f'(x) = 1,$$

а аналитическое определение (1) снова дает

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{x_1 - x}{x_1 - x} = 1,$$

так что

$$\lim \frac{f(x_1) - f(x)}{x_1 - x} = 1 \quad \text{при} \quad x_1 \to x.$$

Простейшим нетривиальным примером является дифференцирование функции

$$y = f(x) = x^2,$$

что в сущности является нахождением наклона параболы. Это — простейший случай, на котором мы можем учиться совершать переход к пределу, когда результат с первого взгляда не очевиден. Мы имеем

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x)}{x_1 - x} = \frac{x_1^2 - x^2}{x_1 - x}.$$

Если бы мы попытались перейти к пределу непосредственно в числителе и в знаменателе, то получили бы не имеющее смысла выражение $\frac{0}{0}$. Но этого затруднения можно избежать, сократив дробь на мешающий нам множитель $x_1 - x$ *до перехода к пределу*. (Такое сокращение законно,

так как при вычислении предела разностного отношения мы считаем, что $x_1 \neq x$, см. стр. 326.) Таким образом мы получаем результат

$$\frac{x_1^2 - x^2}{x_1 - x} = \frac{(x_1 - x)(x_1 + x)}{x_1 - x} = x_1 + x.$$

После сокращения нахождение предела при $x_1 \to x$ не представляет уже никаких трудностей. Этот предел получается путем простой «подстановки», так как разностное отношение в своем новом виде $x_1 + x$ непрерывно, а предел непрерывной функции при $x_1 \to x$ есть просто значение этой функции при $x_1 = x$; в нашем примере мы получаем непосредственно $x + x = 2x$ и, следовательно, если $f(x) = x^2$, то

$$f'(x) = 2x.$$

Совершенно аналогично мы можем доказать, что в случае функции $f(x) = x^3$ мы будем иметь $f'(x) = 3x^2$. В самом деле, отношение

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x)}{x_1 - x} = \frac{x_1^3 - x^3}{x_1 - x}$$

может быть упрощено по формуле

$$x_1^3 - x^3 = (x_1 - x)(x_1^2 + x_1 x + x^2);$$

знаменатель $\Delta x = x_1 - x$ сокращается, и мы получаем непрерывное выражение

$$\frac{\Delta y}{\Delta x} = x_1^2 + x_1 x + x^2.$$

При стремлении x_1 к x это выражение стремится к сумме $x^2 + x^2 + x^2$; в качестве предела получается выражение

$$f'(x) = 3x^2.$$

И вообще, для функции

$$f(x) = x^n,$$

где n — целое положительное, производная будет иметь вид

$$f'(x) = nx^{n-1}.$$

Упражнение. Докажите этот результат. (*Указание*: примените алгебраическую формулу

$$x_1^n - x^n = (x_1 - x)(x_1^{n-1} + x_1^{n-2}x + x_1^{n-3}x^2 + \ldots + x_1 x^{n-2} + x^{n-1}).)$$

В качестве следующего примера, позволяющего непосредственно определить производную, рассмотрим функцию

$$y = f(x) = \frac{1}{x}.$$

Мы имеем

$$\frac{\Delta y}{\Delta x} = \frac{y_1 - y}{x_1 - x} = \left(\frac{1}{x_1} - \frac{1}{x}\right) \cdot \frac{1}{x_1 - x} = \frac{x - x_1}{x_1 x} \cdot \frac{1}{x_1 - x}.$$

Сократим опять дробь и тогда получим

$$\frac{\Delta y}{\Delta x} = -\frac{1}{x_1 x}$$

— выражение, опять-таки непрерывное в точке $x_1 = x$; в качестве предела мы, следовательно, будем иметь

$$f'(x) = -\frac{1}{x^2}.$$

Само собой разумеется, что в данном случае ни производная, ни сама функция не определены в точке $x = 0$.

Упражнение. Докажите аналогичным способом, что функция $f(x) = \frac{1}{x^2}$ имеет производную $f'(x) = -\frac{2}{x^3}$; функция $f(x) = \frac{1}{x^n}$ имеет производную $f'(x) = -\frac{n}{x^{n+1}}$; функция $f(x) = (1+x)^n$ имеет производную $f'(x) = n(1+x)^{n-1}$.

Продифференцируем теперь функцию

$$y = f(x) = \sqrt{x}.$$

В качестве разностного отношения мы получаем

$$\frac{\Delta y}{\Delta x} = \frac{y_1 - y}{x_1 - x} = \frac{\sqrt{x_1} - \sqrt{x}}{x_1 - x}.$$

Воспользовавшись формулой $x_1 - x = (\sqrt{x_1} - \sqrt{x})(\sqrt{x_1} + \sqrt{x})$, можно сократить знаменатель с первым из множителей и получить выражение, непрерывное в точке $x_1 = x$,

$$\frac{\Delta y}{\Delta x} = \frac{1}{\sqrt{x_1} + \sqrt{x}}.$$

Переход к пределу дает

$$f'(x) = \frac{1}{2\sqrt{x}}.$$

Упражнения. Докажите, что функция $f(x) = \frac{1}{\sqrt{x}}$ имеет производную $f'(x) = -\frac{1}{2(\sqrt{x})^3}$; докажите далее:

а) что функция $f(x) = \sqrt[3]{x}$ имеет производную $f'(x) = \frac{1}{3\sqrt[3]{x^2}}$;

б) » » $f(x) = \sqrt{1-x^2}$ » » $f'(x) = -\frac{x}{\sqrt{1-x^2}}$;

в) » » $f(x) = \sqrt[n]{x}$ » » $f'(x) = \frac{1}{n\sqrt[n]{x^{n-1}}}$.

4. Производные от тригонометрических функций. Теперь мы приступим к чрезвычайно важному вопросу — к дифференцированию *тригонометрических функций*. Предварительно условимся, что измерение углов будем производить исключительно в радианах.

Чтобы продифференцировать функцию $y = f(x) = \sin x$, положим $x_1 - x = h$, так что $x_1 = x + h$ и $f(x_1) = \sin x_1 = \sin(x + h)$. Воспользовавшись тригонометрической формулой для синуса суммы двух углов, $\sin(A + B)$, мы получим

$$f(x_1) = \sin(x + h) = \sin x \cos h + \cos x \sin h.$$

Отсюда

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{\sin(x + h) - \sin x}{h} = \cos x \cdot \left(\frac{\sin h}{h}\right) + \sin x \left(\frac{\cos h - 1}{h}\right). \quad (2)$$

Если x_1 стремится к x, то h стремится к 0, $\sin h$ стремится к 0, а $\cos h$ стремится к 1.

Далее, применяя результаты стр. 329–330, мы получим

$$\lim_{h \to 0} \frac{\sin h}{h} = 1 \quad \text{и} \quad \lim_{h \to 0} \frac{\cos h - 1}{h} = 0.$$

Правая часть соотношения (2) стремится, следовательно, к $\cos x$, и мы получаем окончательный результат: *функция* $f(x) = \sin x$ *имеет своей производной функцию* $f'(x) = \cos x$ или, короче,

$$\frac{d(\sin x)}{dx} = \cos x.$$

Упражнение. Докажите, что $\dfrac{d(\cos x)}{dx} = -\sin x$.

Чтобы продифференцировать функцию $f(x) = \operatorname{tg} x$, мы напишем $\operatorname{tg} x = \dfrac{\sin x}{\cos x}$ и получим, далее,

$$\frac{f(x + h) - f(x)}{h} = \left(\frac{\sin(x + h)}{\cos(x + h)} - \frac{\sin x}{\cos x}\right) \cdot \frac{1}{h} =$$
$$= \frac{\sin(x + h)\cos x - \cos(x + h)\sin x}{h} \cdot \frac{1}{\cos(x + h)\cos x} = \frac{\sin h}{h} \cdot \frac{1}{\cos(x + h)\cos x}.$$

(Последнее равенство получается с помощью формулы $\sin(A - B) = \sin A \cos B - \cos A \sin B$, где $A = x + h$, $B = x$.) Если h стремится к 0, то $\dfrac{\sin h}{h}$ стремится к 1, $\cos(x + h)$ стремится к $\cos x$, и отсюда мы делаем заключение:

Производная функции $f(x) = \operatorname{tg} x$ есть функция $f'(x) = \dfrac{1}{\cos^2 x}$, или

$$\frac{d(\operatorname{tg} x)}{dx} = \frac{1}{\cos^2 x}.$$

Упражнение. Докажите, что $\dfrac{d(\operatorname{ctg} x)}{dx} = -\dfrac{1}{\sin^2 x}$.

***5. Дифференцируемость и непрерывность.** *Необходимым условием дифференцируемости некоторой функции является ее непрерывность.* В самом деле, если существует предел дифференциального отношения $\frac{\Delta y}{\Delta x}$ при стремлении Δx к нулю, то ясно, что приращение Δy функции $f(x)$ должно становиться неограниченно малым при стремлении Δx к нулю. Каждая дифференцируемая функция неизбежно является в то же время и непрерывной; поэтому, встречаясь в этой главе не раз с дифференцируемыми функциями, мы воздерживаемся от того, чтобы без особой необходимости постоянно упоминать, что они предполагаются непрерывными, или доказывать их непрерывность.

6. Производная и скорость. Вторая производная и ускорение. До сих пор понятие производной мы связывали с геометрическим представлением графика функции. Однако было бы грубой ошибкой ограничивать роль понятия производной одной лишь задачей об определении наклона касательной к данной кривой. Еще более важной с научной точки зрения задачей является вычисление скорости изменения какой бы то ни было величины $f(t)$, меняющейся с течением времени. Именно с этой стороны Ньютон и подошел к дифференциальному исчислению. В частности, Ньютон стремился проанализировать явление скорости, рассматривая время и положение движущейся частицы как переменные величины (по выражению Ньютона, «флюэнты»). Когда некоторая частица движется вдоль оси x, то ее движение вполне определено, раз задана функция $x = f(t)$, указывающая положение частицы x в любой момент времени t. «Равномерное движение» с постоянной скоростью b по оси x определяется линейной функцией $x = a + bt$, где a есть положение частицы в начальный момент (при $t = 0$).

Движение частицы на плоскости описывается уже двумя функциями

$$x = f(t), \qquad y = g(t),$$

которые определяют ее координаты как функции времени. В частности, равномерному движению соответствуют две линейные функции

$$x = a + bt, \qquad y = c + dt,$$

где b и d — две «компоненты» постоянной скорости, а a и c — координаты начального положения частицы (при $t = 0$); траекторией частицы является прямая линия, уравнение которой

$$(x - a)d - (y - c)b = 0$$

получается путем исключения t из двух стоящих выше соотношений.

Если частица движется в вертикальной плоскости x, y под действием одной лишь силы тяжести, то движение ее (это доказывается в элементарной физике) определено двумя уравнениями

$$x = a + bt, \qquad y = c + dt - \frac{1}{2}\,gt^2,$$

где a, b, c, d — постоянные величины, зависящие от состояния частицы в начальный момент, а g — ускорение силы тяжести, равное приблизительно 9,81, если время измеряется в секундах, а расстояние — в метрах. Траектория движения, получаемая путем исключения t из двух данных уравнений, есть парабола

$$y = c + \frac{d}{b}(x - a) - \frac{1}{2}\, g\, \frac{(x-a)^2}{b^2},$$

если только $b \neq 0$; в противном случае траекторией является отрезок вертикальной оси.

Если частица вынуждена двигаться по некоторой данной кривой (подобно тому как поезд движется по рельсам), то движение ее может быть определено функцией $s(t)$ (функцией времени t), равной длине дуги s, вычисляемой вдоль данной кривой от некоторой начальной точки P_0 до положения частицы в точке P в момент времени t. Например, если речь идет о единичном круге $x^2 + y^2 = 1$, то функция $s = ct$ определяет на этом круге равномерное вращательное движение со скоростью c.

* **Упражнение.** Начертите траектории плоских движений, заданных уравнениями: 1) $x = \sin t$, $y = \cos t$; 2) $x = \sin 2t$, $y = \cos 3t$; 3) $x = \sin 2t$, $y = 2\sin 3t$; 4) в описанном выше параболическом движении предположите начальное положение частицы (при $t = 0$) в начале координат и считайте $b > 0$, $d > 0$. Найдите координаты самой высокой точки траектории. Найдите время t и значение x, соответствующие вторичному пересечению траектории с осью x.

Первой целью, которую поставил себе Ньютон, было нахождение скорости частицы, движущейся неравномерно. Рассмотрим для простоты движение частицы вдоль некоторой прямой линии, заданное функцией $x = f(t)$. Если бы движение было равномерным, т. е. совершалось с постоянной скоростью, то эту скорость можно было бы найти, взяв два момента времени t и t_1 и соответствующие им положения частиц $f(t)$ и $f(t_1)$ и составив отношение

$$v = \text{скорость} = \frac{\text{расстояние}}{\text{время}} = \frac{x_1 - x}{t_1 - t} = \frac{f(t_1) - f(t)}{t_1 - t}. \tag{3}$$

Например, если t измерено в часах, а x в километрах, то при $t_1 - t = 1$ разность $x_1 - x$ будет число километров, пройденных за 1 час, а v — скорость (километров в час). Говоря, что скорость есть величина постоянная, имеют в виду лишь то, что разностное отношение

$$\frac{f(t_1) - f(t)}{t_1 - t} \tag{4}$$

не изменяется при любых значениях t и t_1. Но если движение неравномерно (что имеет, например, место при свободном падении тела, скорость которого по мере падения возрастает), то отношение (4) не дает значения скорости в момент t, а представляет собой то, что принято называть

средней скоростью в промежутке времени от t до t_1. Чтобы получить скорость в *момент* t, нужно вычислить предел средней скорости при стремлении t_1 к t. Таким образом, вместе с Ньютоном определим скорость так:

$$\text{скорость в момент } t = \lim_{t_1 \to t} \frac{f(t_1) - f(t)}{t_1 - t} = f'(t). \tag{5}$$

Другими словами, скорость есть производная от «пройденного пути» (координаты частицы на прямой) по времени, или «мгновенная скорость изменения» пути по отношению ко времени — в противоположность *средней* скорости изменения, определяемой по формуле (4).

Скорость изменения самой скорости называется *ускорением*. Ускорение — это просто производная от производной; она обычно обозначается символом $f''(t)$ и называется *второй производной* от функции $f(t)$.

Галилей заметил, что вертикальное расстояние x, проходимое при свободном падении тела в течение времени t, выражается формулой

$$x = f(t) = \frac{1}{2}\,gt^2, \tag{6}$$

где g есть ускорение силы тяжести. Из формулы (6), путем дифференцирования ее, можно получить скорость v тела в момент времени t; эта скорость выражается формулой

$$v = f'(t) = gt, \tag{7}$$

а ускорение α, которое постоянно,— формулой

$$\alpha = f''(t) = g.$$

Предположим, что нужно найти скорость тела через 2 секунды после начала падения. Найдем сначала среднюю скорость за промежуток времени от $t = 2$ до $t = 2{,}1$:

$$\frac{\frac{1}{2}\,g \cdot (2{,}1)^2 - \frac{1}{2}\,g \cdot 2^2}{2{,}1 - 2} = \frac{4{,}905 \cdot 0{,}41}{0{,}1} = 20{,}11 \text{ (метров в секунду)}.$$

Подставляя же в формулу (7) значение $t = 2$, мы найдем, что значение мгновенной скорости в конце второй секунды равно 19,62 (метров в секунду).

Упражнение. Какова средняя скорость тела за промежуток времени от $t = 2$ до $t = 2{,}01$, от $t = 2$ до $t = 2{,}001$?

При движении точки на плоскости две производные $f'(t)$ и $g'(t)$ двух функций $x = f(t)$ и $y = g(t)$ определяют компоненты скорости. При движении вдоль заданной кривой скорость нужно определить как производную от функции $s = f(t)$, где s — длина дуги.

7. Геометрический смысл второй производной. Вторая производная $f''(x)$ имеет также важное значение в анализе и в геометрии; в самом деле, представляя собой скорость изменения наклона $f'(x)$ кривой $y = f(x)$, вторая производная дает указание на то, как изогнута кривая. Если в некотором промежутке вторая производная больше нуля, то скорость изменения наклона $f'(x)$ положительна. Положительный знак скорости изменения некоторой функции указывает на то, что эта функция возрастает с возрастанием аргумента x. Следовательно, неравенство $f''(x) > 0$ указывает на то, что наклон $f'(x)$ есть возрастающая функция x и, значит, при увеличении x кривая становится более крутой там, где наклон ее положителен, и более пологой там, где наклон отрицателен. Условимся говорить, что в этом случае кривая *вогнута* (рис. 270). Аналогично, если $f''(x) < 0$, то будем говорить, что кривая $y = f(x)$ *выпукла* (рис. 271).

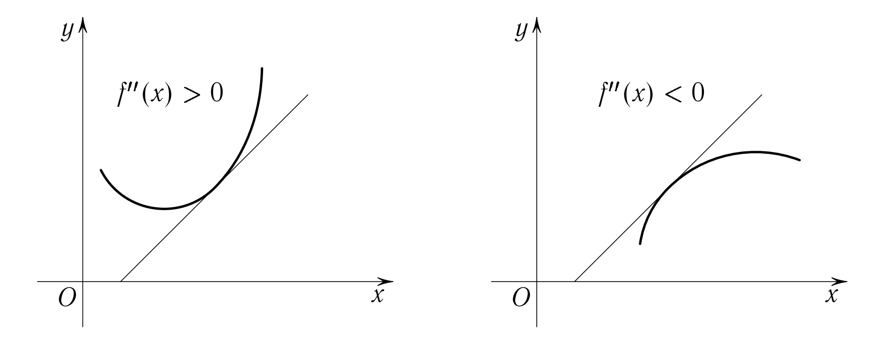


Рис. 270–271. Вогнутость и выпуклость кривой

Парабола $y = f(x) = x^2$ всюду вогнута, так как ее вторая производная $(f''(x) = 2)$ всегда положительна. Кривая $y = f(x) = x^3$ вогнута при $x > 0$ и выпукла при $x < 0$ (рис. 153); это видно по ее второй производной, $f''(x) = 6x$, в чем читатель может легко убедиться сам. Между прочим, при $x = 0$ имеем $f'(x) = 3x^2 = 0$ (но нет ни минимума, ни максимума!), а также $f''(x) = 0$ при $x = 0$. Эта точка называется *точкой перегиба*. В точках, которые так называются, касательная (в данном случае ось x) пересекает кривую.

Если буква s обозначает длину дуги кривой, а буква α — угол наклона, то функция $\alpha = h(s)$ есть функция переменного s. При передвижении точки по кривой функция $\alpha = h(s)$ будет меняться. Скорость этого изменения $h'(s)$ принято называть *кривизной* кривой в точке, для которой длина дуги равна s. Без доказательства отметим, что кривизна k может быть выражена с помощью первой и второй производных от функ-

ции $y = f(x)$, определяющей кривую, согласно следующей формуле:

$$k = \frac{f''(x)}{(1 + (f'(x))^2)^{3/2}}.$$

8. Максимумы и минимумы. Чтобы найти наибольшие и наименьшие значения заданной функции $f(x)$, мы прежде всего должны составить ее производную $f'(x)$, найти затем те значения x, при которых эта производная обращается в нуль, и наконец, исследовать, в каких точках из числа найденных функция имеет максимум и в каких — минимум. Последний из этих вопросов может быть решен с помощью второй производной $f''(x)$, знак которой указывает на выпуклость или вогнутость графика кривой; если же вторая производная обращается в нуль, то обыкновенно это указывает на то, что мы имеем дело с точкой перегиба, и тогда экстремума нет. Принимая во внимание знаки первой и второй производных, можно не только найти экстремумы функции, но и определить вид ее графика. Указанный способ позволяет нам выделить те значения x, при которых функция имеет экстремум; для того чтобы найти соответствующие значения самой функции $y = f(x)$, нужно сделать подстановку найденных значений x в выражение $f(x)$.

В качестве примера рассмотрим многочлен

$$f(x) = 2x^3 - 9x^2 + 12x + 1;$$

его производные выражаются формулами

$$f'(x) = 6x^2 - 18x + 12, \qquad f''(x) = 12x - 18.$$

Квадратное уравнение $f'(x) = 0$ имеет корни $x_1 = 1$, $x_2 = 2$, и в этих точках значения второй производной равны $f''(x_1) = -6 < 0$, $f''(x_2) = 6 > 0$.

Следовательно, функция $f(x)$ имеет максимум $f(x_1) = 6$ и минимум $f(x_2) = 5$.

Упражнения. 1) Наметьте график рассмотренной функции.

2) Исследуйте и наметьте график функции $f(x) = (x^2 - 1)(x^2 - 4)$.

3) Найдите минимум функций $\frac{x+1}{x}$, $x + \frac{a^2}{x}$, $px + \frac{q}{x}$, считая значения p и q положительными. Имеют ли максимум эти функции?

4) Найдите максимум и минимум функций $\sin x$ и $\sin(x^2)$.

§ 3. Техника дифференцирования

До сих пор наши усилия были направлены на то, чтобы продифференцировать целый ряд отдельных функций, причем мы предварительно придавали надлежащий вид разностному отношению. Важным шагом вперед в работах Ньютона, Лейбница и их последователей была замена этих разрозненных индивидуальных приемов мощными общими

методами. С помощью этих методов можно почти автоматически дифференцировать любую функцию из числа тех, с которыми обыкновенно приходится иметь дело в математике; нужно только запастись небольшим числом простых правил и уметь их применять. Совокупность этих приемов приобрела характер вычислительного «алгоритма».

Мы не можем вникать в подробности этой техники. Укажем только немногие наиболее простые правила.

а) *Дифференцирование суммы.* Если a и b — постоянные, и функция $k(x)$ задана формулой

$$k(x) = af(x) + bg(x),$$

то, как это легко докажет сам читатель, справедливо следующее:

$$k'(x) = af'(x) + bg'(x).$$

Аналогичное правило имеет место при любом числе слагаемых.

б) *Дифференцирование произведения.* Производная произведения

$$p(x) = f(x)\,g(x)$$

выражается формулой

$$p'(x) = f(x)\,g'(x) + f'(x)\,g(x).$$

Это легко доказывается следующим приемом: прибавим и отнимем от $p(x+h) - p(x)$ одно и то же выражение, а именно $f(x+h)\,g(x)$:

$$\begin{aligned} p(x+h) - p(x) &= f(x+h)\,g(x+h) - f(x)\,g(x) = \\ &= f(x+h)\,g(x+h) - f(x+h)\,g(x) + f(x+h)\,g(x) - f(x)\,g(x). \end{aligned}$$

Объединяя первые два и последние два члена, мы получим

$$\frac{p(x+h)-p(x)}{h} = f(x+h)\,\frac{g(x+h)-g(x)}{h} + g(x)\,\frac{f(x+h)-f(x)}{h}.$$

Заставим теперь h стремиться к нулю; поскольку $f(x+h)$ при этом стремится к $f(x)$, наше утверждение доказывается немедленно.

Упражнение. Пользуясь этим правилом, докажите, что производная функции $p(x) = x^n$ есть $p'(x) = nx^{n-1}$. (*Указание*: примите во внимание, что $x^n = x \cdot x^{n-1}$, и примените математическую индукцию.)

С помощью правил а) и б) можно дифференцировать любой полином

$$f(x) = a_0 + a_1x + \ldots + a_nx^n:$$

его производная равна выражению

$$f'(x) = a_1 + 2a_2x + 3a_3x^2 + \ldots + na_nx^{n-1}.$$

В качестве одного из применений можно доказать биномиальную теорему (см. стр. 35). Согласно этой теореме, степень бинома $(1+x)^n$ разлагается в полином следующего вида:

$$f(x) = (1+x)^n = 1 + a_1x + a_2x^2 + a_3x^3 + \ldots + a_nx^n, \qquad (1)$$

где коэффициент a_k дается формулой

$$a_k = \frac{n(n-1)\dots(n-k+1)}{k!}. \tag{2}$$

Мы уже видели (упражнение на стр. 444), что дифференцирование левой части формулы (1) дает $n(1+x)^{n-1}$. На основании предыдущего пункта

$$n(1+x)^{n-1} = a_1 + 2a_2x + 3a_3x^2 + \dots + na_nx^{n-1}. \tag{3}$$

Если теперь в этой формуле положить $x = 0$, то получим $n = a_1$, что соответствует формуле (2) при $k = 1$. Снова продифференцируем формулу (3) и тогда будем иметь

$$n(n-1)(1+x)^{n-2} = 2a_2 + 3 \cdot 2a_3x + \dots + n(n-1)a_nx^{n-2}.$$

Подстановка в эту формулу нуля вместо x дает $n(n-1) = 2a_2$, в соответствии с формулой (2) при $k = 2$.

Упражнение. Докажите формулу (2) при $k = 3$ и при любом k (с помощью математической индукции).

в) *Дифференцирование частного.* Если

$$q(x) = \frac{f(x)}{g(x)},$$

то

$$q'(x) = \frac{g(x)\,f'(x) - f(x)\,g'(x)}{(g(x))^2}.$$

Доказательство предоставляется в виде упражнения читателю. Очевидно, нужно предполагать, что $g(x) \neq 0$.

Упражнение. С помощью последнего правила выведите производные от $\operatorname{tg} x$ и $\operatorname{ctg} x$, зная производные от $\sin x$ и $\cos x$. Докажите, что производными от $\sec x = \dfrac{1}{\cos x}$ и $\operatorname{cosec} x = \dfrac{1}{\sin x}$ являются соответственно $\dfrac{\sin x}{\cos^2 x}$ и $-\dfrac{\cos x}{\sin^2 x}$.

Мы умеем теперь дифференцировать любую функцию, являющуюся отношением двух многочленов. Например,

$$f(x) = \frac{1-x}{1+x}$$

имеет производную

$$f'(x) = \frac{-(1+x) - (1-x)}{(1+x)^2} = -\frac{2}{(1+x)^2}.$$

Упражнение. Продифференцируйте функцию

$$f(x) = \frac{1}{x^m} = x^{-m},$$

предполагая m целым положительным. Результат:

$$f'(x) = -mx^{-m-1}.$$

г) *Дифференцирование обратных функций.* Если функции

$$y = f(x) \quad \text{и} \quad x = g(y)$$

взаимно обратны (например, $y = x^2$ и $x = \sqrt{y}$), то производная одной из них является обратной величиной по отношению к производной другой. Именно,

$$g'(y) = \frac{1}{f'(x)}.$$

Это утверждение легко доказать, если обратиться ко взаимно обратным разностным отношениям $\frac{\Delta y}{\Delta x}$ и $\frac{\Delta x}{\Delta y}$; это видно ясно также из геометрической интерпретации обратных функций, приведенной на стр. 302, если отнести наклон касательной к оси y, а не к оси x.

В качестве примера продифференцируем функцию

$$y = f(x) = \sqrt[m]{x} = x^{1/m},$$

обратную по отношению к функции $x = y^m$ (см. также более полное рассуждение, относящееся к случаю $m = \frac{1}{2}$, на стр. 444). Поскольку функция $x = y^m$ имеет своей производной выражение my^{m-1}, то мы имеем

$$f'(x) = \frac{1}{my^{m-1}} = \frac{1}{m} \cdot \frac{y}{y^m} = \frac{1}{m} yy^{-m},$$

откуда, делая подстановки $y = x^{1/m}$ и $y^{-m} = x^{-1}$, получим:

$$f'(x) = \frac{1}{m} x^{1/m-1}.$$

В качестве следующего примера продифференцируем *обратную тригонометрическую функцию* (см. стр. 302)

$$y = \operatorname{arctg} x \quad (\text{что равносильно } x = \operatorname{tg} y).$$

Для того чтобы обеспечить однозначное определение функции y, предположим, что переменная y, обозначающая меру угла в радианах, ограничена промежутком $-\frac{\pi}{2} < y < \frac{\pi}{2}$.

Мы знаем (см. стр. 445), что $\frac{d(\operatorname{tg} y)}{dy} = \frac{1}{\cos^2 y}$, и так как

$$\frac{1}{\cos^2 y} = \frac{\sin^2 y + \cos^2 y}{\cos^2 y} = 1 + \operatorname{tg}^2 y = 1 + x^2,$$

то можно заключить, что

$$\frac{d(\operatorname{arctg} x)}{dx} = \frac{1}{1 + x^2}.$$

Таким же точно путем читатель сможет вывести следующие формулы:

$$\frac{d(\operatorname{arcctg} x)}{dx} = -\frac{1}{1 + x^2},$$

$$\frac{d(\arcsin x)}{dx} = \frac{1}{\sqrt{1 - x^2}},$$

$$\frac{d(\arccos x)}{dx} = -\frac{1}{\sqrt{1 - x^2}}.$$

Наконец, мы приходим к следующему важному правилу:

д) *Дифференцирование сложных функций.* Сложные функции составляются из двух (или нескольких) простых (см. стр. 303). Например функция $z = \sin\sqrt{x}$ составлена из функций $z = \sin y$ и $y = \sqrt{x}$; функция $z = \sqrt{x} + \sqrt{x^5}$ составлена из функций $z = y + y^5$ и $y = \sqrt{x}$; функция $z = \sin(x^2)$ — из функций $z = \sin y$ и $y = x^2$; функция $z = \sin\frac{1}{x}$ — из функций $z = \sin y$ и $y = \frac{1}{x}$.

Если из двух данных функций

$$z = g(y) \quad \text{и} \quad y = f(x)$$

вторую подставить в первую, то получается сложная функция

$$z = k(x) = g[f(x)].$$

Докажем справедливость формулы

$$k'(x) = g'(y)\,f'(x). \tag{4}$$

С этой целью составим разностное отношение

$$\frac{k(x_1) - k(x)}{x_1 - x} = \frac{z_1 - z}{y_1 - y} \cdot \frac{y_1 - y}{x_1 - x},$$

где $y_1 = f(x_1)$ и $z_1 = g(y_1) = k(x_1)$; при стремлении x_1 к x левая часть стремится к $k'(x)$, а два множителя в правой части стремятся соответственно к $g'(y)$ и к $f'(x)$, чем и доказывается формула (4).

В этом доказательстве было необходимо условие $y_1 - y \neq 0$. В самом деле, мы делили на $\Delta y = y_1 - y$; поэтому нужно было считать исключенными те значения x_1, при которых $y_1 - y = 0$. Однако формула (4) остается в силе даже в том случае, если Δy равно нулю в промежутке, окружающем точку x. При этом предположении y остается постоянным, так что $f'(x) = 0$; с другой стороны, и $k(x) = g(y)$ остается постоянным относительно x (поскольку y не меняется при изменении x) и, следовательно, $k'(x) = 0$; итак, формула (4) справедлива также и в рассматриваемом случае.

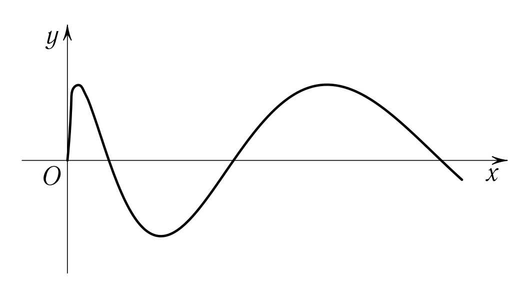


Рис. 272. $y = \sin\sqrt{x}$

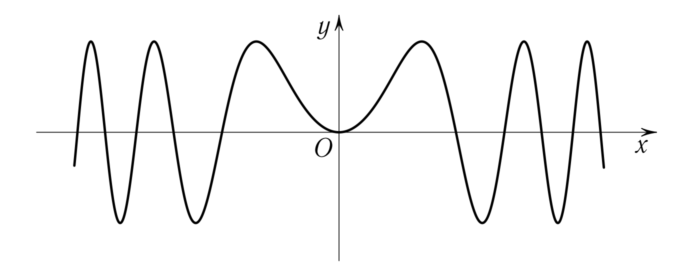


Рис. 273. $y = \sin(x^2)$

Читатель должен проверить следующие примеры:

$$k(x) = \sin\sqrt{x}, \qquad k'(x) = (\cos\sqrt{x})\cdot\frac{1}{2\sqrt{x}};$$

$$k(x) = \sqrt{x} + \sqrt{x^5}, \qquad k'(x) = (1 + 5x^2)\cdot\frac{1}{2\sqrt{x}};$$

$$k(x) = \sin(x^2), \qquad k'(x) = \cos(x^2)\cdot 2x;$$

$$k(x) = \sin\frac{1}{x}, \qquad k'(x) = -\cos\left(\frac{1}{x}\right)\cdot\frac{1}{x^2};$$

$$k(x) = \sqrt{1 - x^2}, \qquad k'(x) = \frac{-1}{2\sqrt{1 - x^2}}\cdot 2x = \frac{-x}{\sqrt{1 - x^2}}.$$

Упражнение. Сопоставляя результаты стр. 451 и стр. 453, докажите, что функция

$$f(x) = \sqrt[m]{x^s}$$

имеет производную

$$f'(x) = \frac{s}{m}x^{s/m-1}.$$

Теперь можно уже отметить, что все наши формулы, касающиеся степеней x, могут быть объединены в одну общую:

При любом положительном или отрицательном рациональном r функция

$$f(x) = x^r$$

имеет производную

$$f'(x) = rx^{r-1}.$$

Упражнения. 1) Произведите дифференцирования в упражнениях на стр. 444, пользуясь только что выведенными правилами.

2) Продифференцируйте следующие функции: $x\sin x$, $\dfrac{\sin nx}{1+x^2}$, $(x^3 - 3x^2 - x + 1)^3$, $1 + \sin^2 x$, $x^2\sin\dfrac{1}{x^2}$, $\arcsin(\cos nx)$, $\operatorname{tg}\dfrac{1+x}{1-x}$, $\operatorname{arctg}\dfrac{1+x}{1-x}$, $\sqrt[4]{1 - x^2}$, $\dfrac{1}{1+x^2}$.

3) Найдите вторые производные от некоторых из вышеприведенных функций и от следующих функций:

$$\frac{1-x}{1+x}, \quad \operatorname{arctg} x, \quad \sin^2 x, \quad \operatorname{tg} x.$$

4) Продифференцируйте функцию

$$u = c_1\sqrt{(x-x_1)^2+y_1^2} + c_2\sqrt{(x-x_2)^2+y_2^2}$$

и докажите минимальные свойства отраженного и преломленного луча, установленные в главе VII (стр. 352 и стр. 403). Предполагается, что отражение и преломление происходят в точке на оси x и что данные начальная и конечная точки заданы координатами (x_1, y_1) и (x_2, y_2).

(*Примечание.* Производная от этой функции обращается в нуль только в одной точке, и поскольку в этой точке неизбежно имеется минимум, а не максимум, нет необходимости исследовать вторую производную.)

Дальнейшие задачи на максимум и минимум

5) Найдите экстремумы следующих функций, наметьте их графики, определите промежутки возрастания, убывания, выпуклости и вогнутости:

$$x^3 - 6x + 2, \quad \frac{1}{1+x^2}, \quad \frac{x^2}{1+x^4}, \quad \cos^2 x.$$

6) Изучите максимумы и минимумы функции $x^3 + 3ax + 1$ в зависимости от значения параметра a.

7) Которая из точек гиперболы $2y^2 - x^2 = 2$ — самая близкая к точке $x = 0$, $y = 3$?

8) Из всех прямоугольников данной площади найдите прямоугольник с самой короткой диагональю.

9) Впишите прямоугольник наибольшей площади в эллипс

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

10) Из всех круговых цилиндров данного объема найдите цилиндр с наименьшей поверхностью.

§ 4. Обозначения Лейбница и «бесконечно малые»

Ньютон и Лейбниц умели находить интегралы и производные как пределы. Но самые основания анализа были долго окружены таинственностью вследствие нежелания признать за понятием предела исключительного права быть источником новых методов. Ни Ньютон, ни Лейбниц не смогли занять ту отчетливую позицию, которая нам кажется простой и естественной теперь, когда понятие предела полностью выяснено. Их пример господствовал больше столетия, в течение которого сущность дела была затемнена бесплодными рассуждениями о «бесконечно малых величинах», о «дифференциалах», о «последнем отношении» и т. д. Неохота, с которой эти понятия были в конце концов отвергнуты, глубоко коренилась в философских концепциях того времени и в самой природе человеческого мышления. Казалось, что можно рассуждать так: конечно, интеграл и производную можно вычислить как пределы. Но все же, в конце концов, чем же являются эти объекты «в себе», независимо от того специфического способа их описания, каким является предельный переход? Ведь интуитивные понятия — такие как площадь или наклон кривой, — имеют как будто бы абсолютный смысл «в себе», и нет надобности в привлечении каких-либо вспомогательных вписанных многоугольников или секущих и их пределов. Без сомнения, желание сформулировать адекватные определения площади или наклона кривой как «вещей в себе» вполне оправдано с психологической точки зрения. Но при зрелых установках, которые так часто расчищали путь к подлинному прогрессу мысли, приходится отбросить это желание и в предельном переходе видеть их единственное приемлемое в научном смысле определение. В XVII в. не было интеллектуальных традиций, которые допускали бы такой радикализм.

Попытка Лейбница «объяснить» производную стоит в непосредственной и безупречной связи с введенным им обозначением для разностного отношения функции $f(x)$

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x)}{x_1 - x}.$$

Предел этого отношения, т. е. производную (которую мы, следуя обычаю, введенному впоследствии Лагранжем, обозначили через $f'(x)$), Лейбниц записывает с помощью символа

$$\frac{dy}{dx},$$

заменяя, таким образом, символ разности «дифференциальным символом» d. Никаких трудностей и никакой таинственности не возникает при условии ясного понимания того, что этот символ является всего лишь указанием на необходимость осуществить предельный переход при

$\Delta x \to 0$, что влечет за собой $\Delta y \to 0$. До перехода к пределу в разностном отношении $\frac{\Delta y}{\Delta x}$ нужно сократить числитель и знаменатель на Δx или суметь преобразовать это отношение так, чтобы переход к пределу совершался безболезненно. Это и является в каждом отдельном случае узловым пунктом процесса дифференцирования. Если бы мы попробовали перейти к пределу без таких предварительных сокращений, то получили бы не имеющее смысла выражение $\frac{\Delta y}{\Delta x} = \frac{0}{0}$, что не принесло бы нам никакой пользы. Таинственность и неясность наступают только в том случае, если мы, по примеру Лейбница или многих его последователей, стали бы говорить нечто подобное следующему: «Δx не стремится к нулю. Напротив, «последнее значение» Δx не есть нуль, а является *«бесконечно малой величиной»*, *«дифференциалом»*, обозначаемым символом dx; аналогично Δy имеет *«последнее»* бесконечно малое значение dy. Настоящее отношение этих бесконечно малых дифференциалов есть опять обыкновенное число $f'(x) = \frac{dy}{dx}$». Поэтому Лейбниц и называл производную «дифференциальным отношением». Такие бесконечно малые величины рассматривались как некие новые числа, хотя и отличные от нуля, но меньшие любого положительного числа из системы действительных чисел.

Считалось, что такие понятия доступны лишь немногим избранным, обладающим настоящим математическим чутьем, и что анализ поэтому, по существу, очень труден, так как не всякий обладает этим чутьем или может его развить. Интеграл, аналогичным образом, рассматривался как сумма «бесконечно большого числа бесконечно малых слагаемых», а именно вида $f(x)\,dx$. Существовало представление, будто такая сумма *есть* интеграл или площадь, в то время как вычисление ее значения как *предела* последовательности *конечных сумм* обыкновенных слагаемых $f(x_j)\Delta x_j$ рассматривалось как некий придаток. Теперь мы попросту отбрасываем желание «непосредственного» объяснения и *определяем* интеграл как предел последовательности конечных сумм. Этим путем все трудности устраняются, и все, что ценно в анализе, приобретает твердую основу.

Несмотря на все сказанное выше, в дальнейшем употребление обозначений Лейбница: $\frac{dy}{dx}$ для производной и $\int f(x)\,dx$ для интеграла, не только сохранилось, но и оказалось чрезвычайно полезным. Против такого употребления нечего возразить, если не упускается из виду, что символ d есть только символ перехода к пределу. Преимущество обозначений Лейбница состоит в том, что с пределами отношений или сумм можно в какой-то мере оперировать так, «как если бы» они были в самом деле отношениями или суммами. Подсказывающая сила этой символики всегда вводила в соблазн приписывать этим символам некоторый совер-

шенно нематематический смысл. Если не поддаваться этому соблазну, то обозначения Лейбница являются по меньшей мере превосходным сокращением более громоздких явных выражений, содержащих предельный переход; по существу же они совершенно необходимы в более развитых частях теории.

Например, правило г) (стр. 453) дифференцирования функции $x = g(y)$, обратной по отношению к функции $y = f(x)$, заключалось в равенстве $g'(y)\, f'(x) = 1$. В обозначениях Лейбница оно выглядит следующим образом: $\frac{dx}{dy} \cdot \frac{dy}{dx} = 1$, т. е. так, «как если бы» можно было сокращать на дифференциалы подобно тому, как это делается с обыкновенными дробями. Аналогично, запись в дифференциальной форме выведенного на стр. 454 правила д) дифференцирования сложной функции $z = k(x)$, где

$$z = g(y), \quad y = f(x),$$

имеет вид

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}.$$

Кроме того, обозначения Лейбница обладают еще и тем преимуществом, что они указывают явно на сами *величины* x, y, z в большей степени, чем на их функциональные взаимоотношения. Последние выражают *процедуру* или совокупность *операций*, с помощью которых из одной величины x получается другая y; например, функция $y = f(x) = x^2$ определяет величину y, равную квадрату величины x. Именно сама операция, в данном случае возведение в квадрат, есть предмет внимания математики. Физики же или инженеры в первую очередь интересуются самими величинами. Поэтому акцентирование самих величин в обозначениях Лейбница имеет особенную привлекательность для лиц, занимающихся прикладной математикой.

Присоединим еще одно замечание. В то время как «дифференциалы» в качестве бесконечно малых величин из математического обихода изгнаны теперь окончательно, и не без позора, само слово «дифференциал» прокралось обратно через заднюю дверь, правда, на этот раз для того, чтобы обозначить безукоризненно законное и полезное понятие. Теперь оно обозначает просто разность Δx, когда Δx мало по сравнению с другими рассматриваемыми величинами. Но здесь мы не можем вступить в обсуждение роли, которую это понятие играет в приближенных вычислениях. Не будем мы также рассматривать другие математические объекты, законно именуемые «дифференциалами» и в некоторых случаях показавшие себя весьма полезными в анализе и его приложениях к геометрии.

§ 5. Основная теорема анализа

1. Основная теорема. Понятие об интегрировании, и в некоторой мере о дифференцировании, было хорошо развито раньше работ Ньютона и Лейбница. Но было совершенно необходимо сделать одно очень простое открытие, для того чтобы дать толчок к огромной эволюции вновь созданного математического анализа. Два как будто бы взаимно не соприкасающихся предельных процесса, употребляемые один для дифференцирования, другой для интегрирования функций, оказались тесно связанными между собой. В самом деле, они являются взаимно обратными операциями, подобно таким операциям, как сложение и вычитание, умножение и деление. Дифференциальное и интегральное исчисления представляют собой нечто единое.

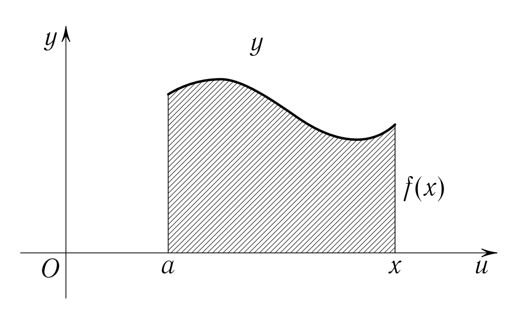


Рис. 274. Интеграл как функция верхнего предела

Великое достижение Ньютона и Лейбница заключается в том, что они впервые ясно осознали и использовали эту *основную теорему анализа*. Без сомнения, их открытие лежало на прямом пути естественного научного развития, и нисколько не удивительно, что различные лица пришли независимо и почти одновременно к ясному пониманию указанного выше обстоятельства.

Для того чтобы точно сформулировать основную теорему, рассмотрим интеграл от функции $y = f(x)$ в пределах от постоянного числа a до числа x, которое будем считать переменным. Чтобы не смешивать верхнего предела интегрирования x с переменной, фигурирующей под знаком интеграла, запишем интеграл в следующем виде (см. стр. 428):

$$F(x) = \int\limits_a^x f(u)\,du, \tag{1}$$

демонстрируя таким образом наше намерение изучать интеграл как функцию $F(x)$ своего верхнего предела (рис. 274). Эта функция $F(x)$ есть площадь под кривой $y = f(u)$ от точки $u = a$ до точки $u = x$. Иногда интеграл $F(x)$ с переменным верхним пределом называют «неопределенным интегралом».

Основная теорема анализа читается следующим образом:

Производная неопределенного интеграла (1) *по его верхнему пределу x равна значению функции $f(u)$ в точке $u = x$:*

$$F'(x) = f(x).$$

Другими словами, процесс интегрирования, ведущий от функции $f(x)$ к функции $F(x)$, «уничтожается» обратным ему процессом дифференцирования, применяемым к функции $F(x)$.

На интуитивной основе доказательство этого предложения не представляет труда. Оно базируется на интерпретации интеграла $F(x)$ как площади, и было бы затемнено, если бы мы попытались представлять функцию $F(x)$ в виде графика и истолковывать производную $F'(x)$ как соответствующий наклон. Оставляя в стороне установленную ранее геометрическую интерпретацию производной, мы сохраним геометрическое толкование интеграла $F(x)$ как площади, а дифференцировать функцию $F(x)$ станем аналитическим методом. Разность

$$F(x_1) - F(x)$$

есть просто площадь под кривой $y = f(u)$ между пределами $u = x_1$ и $u = x$ (рис. 275), и нетрудно понять, что числовое значение этой площади заключено между числами $(x_1 - x)m$ и $(x_1 - x)M$:

$$(x_1 - x)m \leqslant F(x_1) - F(x) \leqslant (x_1 - x)M,$$

где M и m являются, соответственно, наибольшим и наименьшим значениями функции $f(u)$ в промежутке от $u = x$ до $u = x_1$. Действительно, эти произведения дают площади двух прямоугольников, из которых один содержит рассматриваемую криволинейную область, а другой содержится в ней.

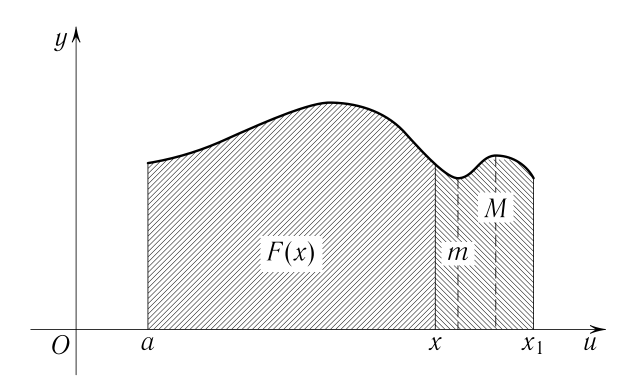


Рис. 275. К доказательству основной теоремы

Отсюда следует

$$m \leqslant \frac{F(x_1) - F(x)}{x_1 - x} \leqslant M.$$

Предположим, что функция $f(u)$ непрерывна, так что при стремлении x_1 к x обе величины M и m стремятся к значению функции $f(u)$ в точке $u = x$, т. е. к значению $f(x)$. В таком случае можно считать

доказанным, что

$$F'(x) = \lim_{x_1 \to x} \frac{F(x_1) - F(x)}{x_1 - x} = f(x). \qquad (2)$$

Интуитивный смысл этого результата заключается в том, что при возрастании скорость изменения площади под кривой $y = f(x)$ равна высоте кривой в точке x.

В некоторых руководствах содержание этой основной теоремы затемняется вследствие неудачно выбранной терминологии. Именно, многие авторы сначала вводят понятие производной, а затем определяют «неопределенный интеграл» просто как результат операции, обратной по отношению к дифференцированию: они говорят, что функция $G(x)$ есть неопределенный интеграл от функции $f(x)$, если

$$G'(x) = f(x).$$

Таким образом, этот способ изложения непосредственно связывает дифференцирование со словом «интеграл». Только позднее вводится понятие «определенный интеграл», трактуемое как площадь или как предел последовательности сумм, причем недостаточно подчеркивается, что слово «интеграл» обозначает теперь нечто совершенно другое, чем прежде. И вот оказывается, что самое главное, что содержится в теории, приобретается лишь украдкой — через заднюю дверь, и учащийся встречается с серьезными затруднениями в своих усилиях понять существо дела. Мы предпочитаем функции $G(x)$, для которых $G'(x) = f(x)$, называть не «неопределенными интегралами», а *первообразными функциями* от функции $f(x)$. Тогда основная теорема может быть сформулирована следующим образом:

Функция $F(x)$, являющаяся интегралом от функции $f(x)$ при постоянном нижнем и переменном верхнем пределе x, есть одна из первообразных функций от функции $f(x)$.

Мы говорим «одна из» первообразных функций по той причине, что если $G(x)$ является первообразной функцией от $f(x)$, то непосредственно ясно, что и любая функция вида $H(x) = G(x) + c$ (c — произвольная постоянная) есть также первообразная, так как $H'(x) = G'(x)$. Обратное утверждение также справедливо. *Две первообразные функции $G(x)$ и $H(x)$ могут отличаться одна от другой не иначе, как постоянным слагаемым.* Действительно, разность $U(x) = G(x) - H(x)$ имеет в качестве производной $U'(x) = G'(x) - H'(x) = f(x) - f(x) = 0$, т. е. эта разность постоянна, так как очевидно, что если график функции в каждой своей точке горизонтален, то сама функция, представляемая графиком, непременно должна быть постоянной.

Это ведет к очень важному правилу вычисления интеграла в пределах от a до b — в предположении, что нам известна какая-либо первообразная функция $G(x)$ от функции $f(x)$. Согласно нашей основной

теореме, функция

$$F(x)=\int_a^x f(u)\,du$$

есть также первообразная функция от функции $f(x)$. Значит, $F(x) = G(x)+c$, где c — постоянная. Значение этой постоянной определится, если мы примем во внимание, что $F(a)=\int_a^a f(u)\,du=0$. Отсюда следует: $0=G(a)+c$, так что $c=-G(a)$. Тогда определенный интеграл в пределах от a до x тождественно удовлетворяет равенству

$$F(x)=\int_a^x f(u)\,du=G(x)-G(a);$$

замена x через b приводит к формуле

$$\int_a^b f(u)\,du=G(b)-G(a), \tag{3}$$

независимо от того, какая именно из первообразных функций была «пущена в ход». Другими словами: *чтобы вычислить определенный интеграл* $\int_a^b f(x)\,dx$*, достаточно найти такую функцию* $G(x)$*, для которой* $G'(x)=f(x)$*, и затем составить разность* $G(b)-G(a)$*.*

2. Первые применения. Интегрирование функций x^r, $\cos x$, $\sin x$. Функция $\operatorname{arctg} x$. Здесь невозможно дать исчерпывающее представление о роли основной теоремы, и мы ограничимся тем, что приведем несколько выразительных примеров. В задачах, встречающихся в механике и физике или в самой математике, очень часто приходится подсчитывать числовое значение некоторого определенного интеграла. Прямая попытка найти интеграл как предел может быть непреодолимо трудной. С другой же стороны, как мы это видели в § 3, любое дифференцирование выполняется сравнительно легко, и без труда возможно накопить очень большое количество формул дифференцирования. Каждая такая формула $G'(x)=f(x)$, обратно, может быть рассматриваема как формула, определяющая первообразную функцию $G(x)$ от функции $f(x)$.

Формула (3) позволяет использовать известную первообразную функцию для вычисления интеграла от функции $f(x)$ в некотором данном промежутке.

Если мы, например, хотим найти интегралы от степеней x^2, x^3, или в общем виде x^n, то самое простое — это действовать, как указано в § 1. По формуле дифференцирования степени производная от x^n равна nx^{n-1},

так что производная от функции

$$G(x) = \frac{x^{n+1}}{n+1} \qquad (n \neq -1)$$

есть функция

$$G'(x) = \frac{n+1}{n+1} x^n = x^n.$$

В таком случае функция $\dfrac{x^{n+1}}{n+1}$ является первообразной функцией по отношению к функции $f(x) = x^n$, а следовательно, мы немедленно получаем формулу

$$\int\limits_a^b x^n\,dx = G(b) - G(a) = \frac{b^{n+1} - a^{n+1}}{n+1}.$$

Это рассуждение несравненно проще громоздкой процедуры непосредственного вычисления интеграла как предела суммы.

Как более общий случай, мы нашли в § 3, что при любом рациональном s, как положительном, так и отрицательном, производная функции x^s равна sx^{s-1}, а потому при $s = r + 1$ функция

$$G(x) = \frac{x^{r+1}}{r+1}$$

имеет производную $f(x) = G'(x) = x^r$ (мы предполагаем, что $r \neq -1$, т. е. что $s \neq 0$). Итак, функция $\dfrac{x^{r+1}}{r+1}$ есть первообразная функция, или «неопределенный интеграл» от x^r, и мы получаем (при положительных a и b и при $r \neq -1$) формулу

$$\int\limits_a^b x^r\,dx = \frac{b^{r+1} - a^{r+1}}{r+1}. \tag{4}$$

В формуле (4) приходится предполагать, что стоящая под интегралом функция x^r определена и непрерывна в промежутке интегрирования, так что нужно исключить точку $x = 0$, если $r < 0$. Вот потому мы и вынуждены допустить, что в этом случае a и b положительны.

Если положим $G(x) = -\cos x$, то получим $G'(x) = \sin x$, и отсюда возникает соотношение

$$\int\limits_0^a \sin x\,dx = -(\cos a - \cos 0) = 1 - \cos a.$$

Аналогично, если $G(x) = \sin x$, то $G'(x) = \cos x$, и значит,

$$\int\limits_0^a \cos x\,dx = \sin a - \sin 0 = \sin a.$$

Особенно интересный результат получается из формулы дифференцирования функции $\operatorname{arctg} x$:

$$\frac{d(\operatorname{arctg} x)}{dx} = \frac{1}{1+x^2}.$$

Раз функция $\operatorname{arctg} x$ есть первообразная по отношению к функции $\frac{1}{1+x^2}$, то на основании формулы (3) можно написать

$$\operatorname{arctg} b - \operatorname{arctg} 0 = \int_0^b \frac{1}{1+x^2}\,dx.$$

Но $\operatorname{arctg} 0 = 0$ (нулевому значению тангенса соответствует нулевое значение угла). Итак, мы имеем

$$\operatorname{arctg} b = \int_0^b \frac{1}{1+x^2}\,dx. \tag{5}$$

В частности, если $b = 1$, то $\operatorname{arctg} b$ равно $\frac{\pi}{4}$ $\Big($значению тангенса, равному 1, соответствует угол в 45°, что в радианной мере составляет $\frac{\pi}{4}\Big)$. Таким образом, мы получаем замечательную формулу

$$\frac{\pi}{4} = \int_0^1 \frac{1}{1+x^2}\,dx. \tag{6}$$

Это показывает, что площадь под графиком функции $y = \frac{1}{1+x^2}$ в пределах от $x = 0$ до $x = 1$ равна четверти площади единичного круга.

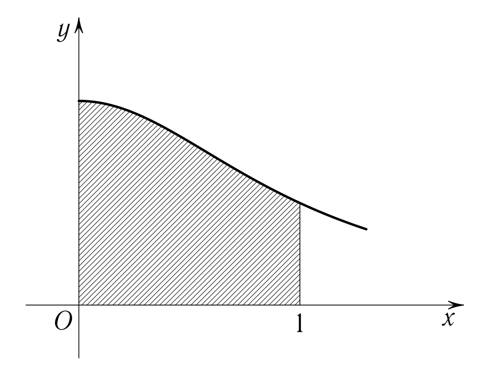


Рис. 276. Площадь под кривой $y = \frac{1}{1+x^2}$ в пределах от 0 до 1 равна $\frac{\pi}{4}$

3. Формула Лейбница для π. Последний результат приводит к одной из красивейших математических формул, открытых в XVII в., — к знакопеременному ряду Лейбница, позволяющему вычислять π:

$$\frac{\pi}{4} = \frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \ldots \tag{7}$$

Символ $+\ldots$ следует понимать в том смысле, что последовательность конечных «частных сумм», получающихся, когда в правой части равенств берется лишь n членов суммы, стремится к пределу $\frac{\pi}{4}$ при неограниченном возрастании n.

Чтобы доказать эту замечательную формулу, нам достаточно вспомнить формулу суммы конечной геометрической прогрессии

$$\frac{1-q^n}{1-q} = 1 + q + q^2 + \ldots + q^{n-1},$$

или

$$\frac{1}{1-q} = 1 + q + q^2 + \ldots + q^{n-1} + \frac{q^n}{1-q}.$$

Если в последнее алгебраическое тождество подставим $q = -x^2$, то получим

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \ldots + (-1)^{n-1}x^{2n-2} + R_n, \tag{8}$$

где «остаточный член» R_n выражается формулой

$$R_n = (-1)^n \frac{x^{2n}}{1+x^2}.$$

Равенство (8) можно проинтегрировать в пределах от 0 до 1. Следуя правилу а) из § 3, мы должны взять в правой части сумму интегралов от отдельных слагаемых. На основании (4) мы знаем, что

$$\int\limits_a^b x^m\,dx = \frac{b^{m+1} - a^{m+1}}{m+1},$$

откуда, в частности, получим $\int\limits_0^1 x^m\,dx = \frac{1}{m+1}$, а следовательно,

$$\int\limits_0^1 \frac{dx}{1+x^2} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots + (-1)^{n-1}\frac{1}{2n-1} + T_n, \tag{9}$$

где $T_n = (-1)^n \cdot \int\limits_0^1 \frac{x^{2n}}{1+x^2}\,dx$. Согласно формуле (5), левая часть формулы (9) равна $\frac{\pi}{4}$. Разность между $\frac{\pi}{4}$ и частной суммой

$$S_n = 1 - \frac{1}{3} + \frac{1}{5} - \ldots + \frac{(-1)^{n-1}}{2n-1}$$

равна $\frac{\pi}{4} - S_n = T_n$. Остается доказать, что T_n стремится к нулю при возрастании n. Мы имеем неравенство

$$\frac{x^{2n}}{1+x^2} \leqslant x^{2n}.$$

Вспомнив формулу (13) § 1, устанавливающую неравенство

$$\int\limits_a^b f(x)\,dx \leqslant \int\limits_a^b g(x)\,dx \quad \text{при} \quad f(x) \leqslant g(x) \quad \text{и} \quad a < b,$$

мы видим, что

$$|T_n| = \int_0^1 \frac{x^{2n}}{1+x^2}\,dx \leqslant \int_0^1 x^{2n}\,dx.$$

Правая часть в этом неравенстве, согласно формуле (4), равна $\frac{1}{2n+1}$; поэтому $|T_n| < \frac{1}{2n+1}$. Окончательно имеем неравенство

$$\left|\frac{\pi}{4} - S_n\right| < \frac{1}{2n+1}.$$

Так как $\frac{1}{2n+1}$ стремится к нулю, то это и показывает, что S_n стремится к $\frac{\pi}{4}$ при возрастании n. Таким образом, формула Лейбница доказана.

§ 6. Показательная (экспоненциальная) функция и логарифм

Концепции анализа предоставляют возможность построить гораздо более полную теорию логарифма и показательной функции, чем это делает та элементарная процедура, которая лежит в основе обычного преподавания в школе. Там обычно отправляются от целых степеней a^n положительного числа a, а затем определяют корень $a^{1/m} = \sqrt[m]{a}$, получая, таким образом, значения a^r при любом рациональном показателе $r = \frac{n}{m}$. Затем значение степени a^x при иррациональном x определяется так, что a^x должна быть непрерывной функцией от x,— деликатный вопрос, обыкновенно опускаемый в элементарном изложении. Наконец, логарифмом числа y при основании a называется функция, обратная по отношению к показательной функции $y = a^x$.

В последующем изложении теории этих функций, построенном на основах анализа, ход мыслей противоположный. Мы начнем с логарифма, а затем придем к показательной функции.

1. Определение и свойства логарифма. Эйлерово число e. Определим логарифм, или, точнее говоря, «натуральный логарифм» $F(x) = \ln x$ (его связь с обычным десятичным логарифмом будет установлена в пункте 2), как площадь под кривой $y = \frac{1}{u}$ в пределах от $u = 1$ до $u = x$, или, что сводится к тому же, как следующий интеграл:

$$F(x) = \ln x = \int_1^x \frac{1}{u}\,du \tag{1}$$

(см. рис. 5, стр. 47). Здесь переменная x может быть любым положительным числом. Нуль исключается потому, что при стремлении u к нулю функция, стоящая под интегралом, стремится к бесконечности.

Естественно заняться изучением функции $F(x)$. Мы знаем, что первообразная функция по отношению к любой степени x^n представляет собой функцию того же типа, так как равна $\dfrac{x^{n+1}}{n+1}$; исключением является степень $n=-1$. В этом последнем случае знаменатель $n+1$ превратился бы в нуль, и формула (4) на стр. 464 потеряла бы смысл. Таким образом, можно ожидать, что изучение интеграла от функции $\dfrac{1}{x}$ или $\dfrac{1}{u}$ приведет к новому (и интересному) типу функции.

Хотя мы и принимаем формулу (1) за определение функции $\ln x$, однако мы не «знаем» самой функции, пока мы не установим ее свойств и не найдем способов находить ее числовые значения. Нужно заметить, что для современных методов в анализе очень характерно то, что мы отправляемся от общих понятий — таких как площадь или интеграл, и уже на основе этих понятий устанавливаем определения, подобные (1); затем выводим свойства определяемых объектов и лишь в самом конце приходим к явным выражениям, позволяющим вычислять их числовые значения.

Первое важное свойство функции $\ln x$ непосредственно следует из основной теоремы § 5. Согласно этой теореме, справедливо равенство

$$F'(x)=\frac{1}{x}. \tag{2}$$

Из формулы (2) следует, что производная $F(x)$ всегда положительна, а это указывает, очевидно, на то, что функция $\ln x$ монотонно возрастает при возрастании x.

Главное свойство логарифма выражается формулой

$$\ln a+\ln b=\ln(ab). \tag{3}$$

Значение этой формулы в практических применениях логарифмов к числовым выкладкам хорошо известно. Формулу (3) можно было бы получить интуитивно, воспользовавшись площадями, определяющими три величины, а именно: $\ln a$, $\ln b$ и $\ln(ab)$. Но мы предпочтем развернуть доказательство, типичное для анализа: наряду с функцией $F(x)=\ln x$ рассмотрим другую функцию

$$k(x)=\ln(ax)=\ln w=F(w),$$

полагая $w=f(x)=ax$, где a — произвольная положительная постоянная. Функцию $k(x)$ можно легко продифференцировать с помощью правила д) из § 3: $k'(x)=f'(w)\cdot f'(x)$. Вследствие формулы (2) и поскольку $f'(x)=a$, это выражение принимает вид

$$k'(x)=\frac{a}{w}=\frac{a}{ax}=\frac{1}{x}.$$

Итак, функция $k(x)$ имеет ту же производную, что и функция $F(x)$; раз так, то, согласно сказанному на стр. 462, мы имеем тождество

$$\ln(ax)=k(x)=F(x)+c,$$

где c есть постоянная, не зависящая от значения переменной x. Константа c определяется с помощью простой подстановки $x = 1$ в последнем равенстве. Из определения (1) следует, что

$$F(1) = \ln 1 = 0$$

(так как интеграл, взятый в качестве определения, при значении $x = 1$ имеет равные верхний и нижний пределы). Теперь мы можем написать

$$k(1) = \ln(a \cdot 1) = \ln a = \ln 1 + c = c,$$

т. е. $c = \ln a$, а потому при любом x справедливо тождество

$$\ln(ax) = \ln a + \ln x. \tag{3a}$$

Полагая $x = b$, мы получим, наконец, искомую формулу (3).

В частности, при $a = x$ мы найдем последовательно, что

$$\left.\begin{aligned} \ln(x^2) &= 2\ln x, \\ \ln(x^3) &= 3\ln x, \\ &\dots\dots\dots \\ \ln(x^n) &= n\ln x. \end{aligned}\right\} \tag{4}$$

Из равенств (4) можно заключить, что при неограниченном возрастании x значения функции $\ln x$ также возрастают неограниченно. Достаточно заметить, например, что

$$\ln(2^n) = n\ln 2,$$

причем правая часть, очевидно, неограниченно возрастает вместе с n, и вспомнить, что было установлено свойство монотонного возрастания функции $\ln x$. Далее, мы имеем:

$$0 = \ln 1 = \ln\left(x \cdot \frac{1}{x}\right) = \ln x + \ln\frac{1}{x},$$

так что

$$\ln\frac{1}{x} = -\ln x. \tag{5}$$

Наконец, справедливо равенство

$$\ln x^r = r\ln x \tag{6}$$

при любом рациональном показателе $r = \dfrac{m}{n}$. В самом деле, полагая $x^r = u$, мы получаем:

$$n\ln u = \ln u^n = \ln x^{\frac{m}{n}\cdot n} = \ln x^m = m\ln x,$$

откуда следует

$$\ln x^{\frac{m}{n}} = \frac{m}{n}\ln x.$$

Поскольку $\ln x$ есть монотонная и непрерывная функция от x, принимающая значение 0 при $x = 1$ и стремящаяся к бесконечности при

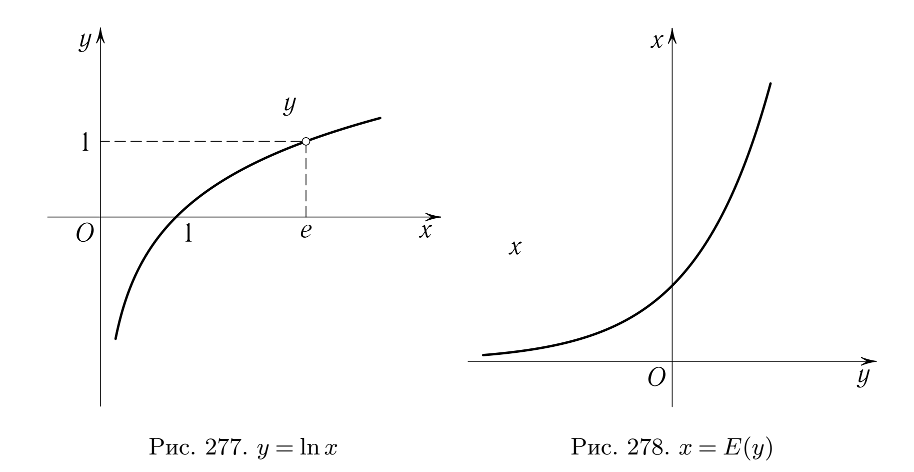


Рис. 277. $y = \ln x$

Рис. 278. $x = E(y)$

неограниченном возрастании x, должно существовать некоторое число x, большее чем единица и такое, что для него будет иметь место равенство $\ln x = 1$. Следуя Эйлеру, обозначим это число буквой e. (Тождественность этого определения с определением, данным на стр. 320, будет доказана позднее.) Итак, число e определено уравнением

$$\ln e = 1. \tag{7}$$

Мы ввели число e, опираясь на свойство непрерывных функций, обеспечивающее существование корня этого уравнения. Теперь мы продолжим наше изыскание, чтобы как *следствие* получить явные формулы, позволяющие вычислить e с какой угодно точностью.

2. Показательная (экспоненциальная) функция. Суммируя наши предыдущее результаты, мы можем сказать, что функция $F(x) = \ln x$ равна нулю при $x = 1$; монотонно возрастает до бесконечности при $x \to \infty$ $\Big($но при этом график имеет убывающий наклон, равный величине $\dfrac{1}{x}\Big)$; при значениях x, меньших единицы, выражается при помощи функции $-\ln\dfrac{1}{x}$, так что $\ln x$ стремится к отрицательной бесконечности при $x \to 0$.

Монотонный характер возрастания функции $y = \ln x$ позволяет рассматривать обратную ей функцию

$$x = E(y),$$

график которой (рис. 278) получается обычным путем из графика функции $y = \ln x$ (рис. 277); эта обратная функция определена при всех значениях y от $-\infty$ до $+\infty$. При $y \to -\infty$ функция $E(y)$ стремится к нулю; при $y \to \infty$, с другой стороны, $E(y) \to \infty$. Рассматриваемая функция E

обладает следующим основным свойством:

$$E(a) \cdot E(b) = E(a + b) \tag{8}$$

при любой паре значений a и b. Последнее тождество есть просто видоизменение формулы (3), выражающей свойство логарифма. Действительно, если мы придадим формуле (3) вид $\ln x + \ln z = \ln(xz)$ и затем положим

$$E(a) = x, \quad E(b) = z \quad (\text{т. е. } a = \ln x, \;\; b = \ln z),$$

то будем иметь

$$\ln(xz) = \ln x + \ln z = a + b,$$

а отсюда вытекает

$$E(a + b) = xz = E(a) \cdot E(b),$$

что и требовалось доказать.

Так как, по определению, $\ln e = 1$, то имеет место соотношение

$$E(1) = e;$$

присоединяя к этому формулу (8), получим равенство $e^2 = E(1) \cdot E(1) = E(2)$, и т. д. Вообще,

$$E(n) = e^n$$

при любом целом n. Аналогично можно получить $E\left(\frac{1}{n}\right) = e^{\frac{1}{n}}$, так что

$$E\left(\frac{p}{q}\right) = E\left(\frac{1}{q}\right) \cdot \ldots \cdot E\left(\frac{1}{q}\right) = (e^{\frac{1}{q}})^p;$$

полагая затем $\frac{p}{q} = r$, заключаем, что

$$E(r) = e^r$$

при любом рациональном r. Поэтому вполне естественно определить иррациональную степень числа e по формуле

$$e^y = E(y),$$

справедливой при любом действительном y, поскольку функция E непрерывна при всех значениях y и тождественна с функцией e^y при рациональных значениях y. Формулу (8), выражающую основное свойство функции E, или, по общепринятой терминологии, экспоненциальной (показательной) функции, теперь можно выразить при помощи равенства

$$e^a e^b = e^{a+b}, \tag{9}$$

которое тем самым установлено для произвольных рациональных или иррациональных значений a и b.

Во всех этих рассуждениях мы относили логарифм и показательную функцию к числу e как к «основанию», точнее, к «натуральному

основанию» логарифмов. Перейти от основания e к некоторому другому положительному основанию не представляет труда. Начнем с рассмотрения *натурального* логарифма

$$\alpha = \ln a$$

(что равносильно $a = a^{\alpha} = e^{\ln a}$). Показательную функцию a^x мы станем *определять* посредством следующего сложного выражения:

$$z = a^x = e^{\alpha x} = e^{x \ln a}. \tag{10}$$

Например,

$$10^x = e^{x \ln 10}.$$

Назовем функцию, обратную по отношению к функции a^x, *логарифмом при основании* a; нетрудно понять, что *натуральный логарифм* от z есть произведение x на α: другими словами, логарифм числа z при основании a получается путем деления натурального логарифма числа z на постоянный натуральный логарифм числа a. Если $a = 10$, то это число (с четырьмя значащими цифрами) выражается следующим образом:

$$\ln 10 \approx 2{,}303.$$

3. Формулы дифференцирования функций e^x, a^x, x^s. Так как показательную функцию мы определили как обратную по отношению к функции $y = \ln x$, то из правила дифференцирования обратных функций (§ 3) вытекает, что

$$E'(y) = \frac{dx}{dy} = \frac{1}{\dfrac{dy}{dx}} = \frac{1}{\dfrac{1}{x}} = x = E(y),$$

т. е.

$$E'(y) = E(y). \tag{11}$$

Производная от «натуральной» показательной функции тождественно равна самой функции. Это есть истинный источник всех свойств показательной функции и основная причина ее роли во всех приложениях, как это станет видно в последующих разделах. Используя дифференциальные обозначения, мы можем записать формулу (11) в следующем виде:

$$\frac{d}{dx} e^x = e^x. \tag{11a}$$

В более общем случае, дифференцируя сложную функцию

$$f(x) = e^{\alpha x}$$

с помощью правила, данного в § 3, мы получим равенство

$$f'(x) = \alpha e^{\alpha x} = \alpha f(x).$$

Таким образом, полагая $\alpha = \ln a$, мы найдем, что функция

$$f(x) = a^x$$

имеет производную

$$f'(x) = a^x \ln a.$$

Переходя теперь к рассмотрению степенной функции

$$f(x) = x^s$$

при любом действительном показателе s и при положительном переменном x, мы можем определить ее по формуле

$$x^s = e^{s \ln x}.$$

Снова применяя правило дифференцирования сложной функции к случаю, когда $f(x) = e^{sz}$, $z = \ln x$, мы найдем производную

$$f'(x) = se^{sz} \cdot \frac{1}{x} = sx^s \cdot \frac{1}{x},$$

так что

$$f'(x) = sx^{s-1},$$

в полном соответствии с прежним правилом дифференцирования степенной функции при рациональном показателе s.

4. Явные выражения числа e и функций e^x и $\ln x$ в виде пределов. Для того чтобы найти явные формулы, выражающие эти функции, мы используем формулы дифференцирования показательной и логарифмической функции. Так как производная функции $\ln x$ равна $\frac{1}{x}$, то в силу определения производной мы получаем соотношение

$$\frac{1}{x} = \lim \frac{\ln x_1 - \ln x}{x_1 - x} \quad \text{при} \quad x_1 \to x.$$

Положим $x_1 = x + h$ и допустим, что h стремится к нулю, пробегая последовательность $h = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots, \frac{1}{n}, \ldots$; тогда, применяя правила действий с логарифмами, мы получим

$$\frac{\ln\left(x + \frac{1}{n}\right) - \ln x}{\frac{1}{n}} = n \ln \frac{x + \frac{1}{n}}{x} = \ln\left[\left(1 + \frac{1}{nx}\right)^n\right] \to \frac{1}{x}.$$

Если вместо $\frac{1}{x}$ мы подставим z и перейдем к пределу, то написанное выше соотношение примет вид

$$z = \lim \ln\left[\left(1 + \frac{z}{n}\right)^n\right] \quad \text{при} \quad n \to \infty.$$

Или, в терминах показательной функции,

$$e^z = \lim \left(1 + \frac{z}{n}\right)^n \quad \text{при} \quad n \to \infty. \qquad (12)$$

Мы получили общеизвестную формулу, определяющую показательную функцию просто как предел. В частности, при $z = 1$ эта формула дает

$$e = \lim \left(1 + \frac{1}{n}\right)^n \quad \text{при} \quad n \to \infty, \qquad (13)$$

а при $z = -1$ получается

$$\frac{1}{e} = \lim \left(1 - \frac{1}{n}\right)^n \quad \text{при} \quad n \to \infty. \tag{13a}$$

Эти выражения сразу ведут к разложениям в бесконечные ряды. По биномиальной теореме можно написать

$$\left(1 + \frac{x}{n}\right)^n = 1 + n\frac{x}{n} + \frac{n(n-1)}{2!}\frac{x^2}{n^2} + \frac{n(n-1)(n-2)}{3!}\frac{x^3}{n^3} + \ldots + \frac{x^n}{n^n},$$

или

$$\left(1 + \frac{x}{n}\right)^n = 1 + \frac{x}{1!} + \frac{x^2}{2!}\left(1 - \frac{1}{n}\right) + \frac{x^3}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \ldots$$
$$\ldots + \frac{x^n}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\ldots\left(1 - \frac{n-2}{n}\right)\left(1 - \frac{n-1}{n}\right).$$

Позволительно догадываться и нетрудно полностью доказать (подробности мы здесь опускаем), что можно перейти к пределу при $n \to \infty$ путем замены в каждом члене величины $\frac{1}{n}$ на 0. Это дает хорошо известный бесконечный ряд, который служит для вычисления функции e^x:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \tag{14}$$

и, в частности, при $x = 1$ ряд, сходящийся к пределу e:

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots,$$

чем устанавливается идентичность e с тем числом, определение которого дано на стр. 320. При $x = -1$ получается ряд

$$\frac{1}{e} = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \ldots,$$

который дает превосходное приближение уже при очень малом числе членов, так как ошибка, которую мы совершаем, обрывая ряд на n-м члене, меньше величины $(n+1)$-го члена.

Пользуясь формулой дифференцирования показательной функции, можно получить интересное выражение для логарифма. Имеет место соотношение

$$\lim_{h \to 0} \frac{e^h - 1}{h} = \lim_{h \to 0} \frac{e^h - e^0}{h} = 1,$$

так как указанный предел есть не что иное, как значение производной от функции e^y при $y = 0$, а таковое равно 1. Подставим в эту формулу вместо h значение $\frac{z}{n}$, где z — произвольное число, а n пусть пробегает последовательность целых положительных чисел. Тогда мы получим

$$n\,\frac{e^{\frac{z}{n}} - 1}{z} \to 1,$$

или

$$n(\sqrt[n]{e^z} - 1) \to z$$

при $n \to \infty$. Вводя обозначение $z = \ln x$, или $e^z = x$, можно окончательно написать

$$\ln x = \lim n(\sqrt[n]{x} - 1) \quad \text{при} \quad n \to \infty. \tag{15}$$

Поскольку $\sqrt[n]{x} \to 1$ при $n \to \infty$ (см. стр. 346), формула (15) представляет логарифм в виде произведения двух множителей, из которых первый стремится к бесконечности, а второй — к нулю.

Примеры и упражнения

Введение показательной и логарифмической функций доставляет возможность оперировать с функциями достаточно обширного класса и открывает доступ ко многим приложениям.

Продифференцируйте: 1) $x(\ln x - 1)$; 2) $\ln(\ln x)$; 3) $\ln(x + \sqrt{1 + x^2})$; 4) $\ln(x + \sqrt{1 - x^2})$; 5) e^{-x^2}; 6) e^{e^x} (сложная функция e^z, где $z = e^x$); 7) x^x (*Указание*: $x^x = e^{x \ln x}$); 8) $\ln \operatorname{tg} x$; 9) $\ln \sin x$, $\ln \cos x$; 10) $\dfrac{x}{\ln x}$.

Найдите максимумы и минимумы функций: 11) xe^{-x}; 12) $x^2 e^{-x}$; 13) xe^{-ax}.

*14) Найдите геометрическое место максимумов кривой $y = xe^{-ax}$ при переменном параметре a.

15) Покажите, что все последовательные производные от функции e^{-x^2} имеют вид произведения множителя e^{-x^2} на многочлены от x последовательно возрастающих степеней.

*16) Покажите, что производные n-го порядка от функции $e^{-\frac{1}{x^2}}$ имеют вид произведения множителей $e^{-\frac{1}{x^2}} \cdot \dfrac{1}{x^{3n}}$ на многочлен степени $2n - 2$.

*17) *Логарифмическое дифференцирование.* Дифференцирование произведений может быть иногда упрощено путем применения основного свойства логарифма. Действительно, имея дело с произведением

$$p(x) = f_1(x) f_2(x) \ldots f_n(x),$$

можно написать

$$\frac{d(\ln p(x))}{dx} = \frac{d(\ln f_1(x))}{dx} + \frac{d(\ln f_2(x))}{dx} + \ldots + \frac{d(\ln f_n(x))}{dx},$$

и дальше, с помощью формулы дифференцирования сложных функций, отсюда следует

$$\frac{p'(x)}{p(x)} = \frac{f_1'(x)}{f_1(x)} + \frac{f_2'(x)}{f_2(x)} + \ldots + \frac{f_n'(x)}{f_n(x)}.$$

Воспользуйтесь этим при дифференцировании примеров

а) $x(x + 1)(x + 2) \ldots (x + n)$; б) xe^{-ax^2}.

5. Бесконечный ряд для логарифма. Вычисление логарифмов. Числовые значения логарифмов вычисляются отнюдь не с помощью формулы (15). Гораздо лучше приспособлено для этой цели совершенно иное, более полезное, явное выражение, имеющее, кроме того, большое теоретическое значение. С помощью метода, примененного на стр. 466 при вычислении π, мы получим это выражение, опираясь на определение логарифма по формуле (1). Но здесь необходим один

предварительный шаг: вместо функции $\ln x$ рассмотрим функцию $y = \ln(1+x)$, составленную из функций $y = \ln z$ и $z = 1 + x$. Имеем:

$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx} = \frac{1}{z} \cdot 1 = \frac{1}{1+x}.$$

Итак, функция $y = \ln(1+x)$ является первообразной по отношению к функции $\frac{1}{1+x}$, и мы заключаем, согласно основной теореме, что интеграл от функции $\frac{1}{1+u}$ в пределах от 0 до x равен выражению $\ln(1+x) - \ln 1 = \ln(1+x)$; или, в символической записи,

$$\ln(1+x) = \int\limits_0^x \frac{1}{1+u}\,du. \tag{16}$$

(Эту формулу можно было бы, конечно, получить и интуитивно из геометрической интерпретации логарифма как площади. Сравните с рассуждением на стр. 468.)

В формулу (16) подставим вместо $(1+u)^{-1}$ сумму геометрической прогрессии, как мы это делали на стр. 466, а именно

$$\frac{1}{1+u} = 1 - u + u^2 - u^3 + \ldots + (-1)^{n-1}u^{n-1} + (-1)^n \frac{u^n}{1+u};$$

из осторожности мы предпочитаем оперировать не с бесконечным рядом, а с конечной суммой и остаточным членом, который равен

$$R_n = (-1)^n \frac{u^n}{1+u}.$$

Подставив эту сумму в формулу (16), можно применить правило почленного интегрирования конечной суммы. Интеграл от степени u^s в пределах от 0 до x равен $\frac{x^{s+1}}{s+1}$; таким образом, мы получим немедленно

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^{n-1}\frac{x^n}{n} + T_n,$$

где остаточный член T_n выражается интегралом

$$T_n = (-1)^n \int\limits_0^x \frac{u^n}{1+u}\,du.$$

Покажем теперь, что T_n стремится к нулю при возрастании n, предварительно условившись, что переменное x может быть лишь больше -1 и не превышать $+1$, т. е., другими словами, что выполнено неравенство

$$-1 < x \leqslant 1$$

(заметим, что $x = +1$ включается, в то время как $x = -1$ не включается). Согласно нашему предположению, в промежутке интегрирования переменное u больше, чем некоторое число $-\alpha$, которое может быть

близко к -1, но во всяком случае больше, чем -1, так что $-1 < -\alpha < u$. Отсюда следует $0 < 1 - \alpha < 1 + u$. Поэтому при условии, что u заключено в промежутке от 0 до x, имеет место неравенство

$$\left|\frac{u^n}{1+u}\right| \leqslant \frac{|u|^n}{1-\alpha},$$

и следовательно,

$$|T_n| \leqslant \frac{1}{1-\alpha}\left|\int\limits_0^x u^n\,du\right|,$$

или

$$|T_n| \leqslant \frac{1}{1-\alpha}\,\frac{|x|^{n+1}}{n+1} \leqslant \frac{1}{1-\alpha}\,\frac{1}{n+1}.$$

Поскольку число $1-\alpha$ является постоянным, мы видим, что выражение, стоящее справа, а следовательно, и стоящее слева $|T_n|$, при возрастании n стремится к нулю; значит, из неравенства

$$\left|\ln(1+x) - \left(x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots + (-1)^n\frac{x^n}{n}\right)\right| \leqslant \frac{1}{1-\alpha}\,\frac{1}{n+1} \qquad (17)$$

вытекает, что при $-1 < x \leqslant 1$ справедливо равенство

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots \qquad (18)$$

Подставляя, в частности, $x = 1$, получаем любопытную формулу

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots \qquad (19)$$

Эта формула по своей структуре похожа на выведенную раньше формулу, представляющую в виде ряда число $\frac{\pi}{4}$.

Ряд (18) не имеет большого практического значения для вычисления логарифмов, потому что область изменения величины $1 + x$ ограничена промежутком от 0 до 2, а также по той причине, что сходимость ряда очень медленная: пришлось бы брать много членов, чтобы получить сколько-нибудь точный результат. При помощи следующего приема мы получим выражение, практически более удобное. Вместо x в формулу (18) подставим $-x$:

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \ldots \qquad (20)$$

Вычитая, далее, формулу (20) из формулы (18) и применяя преобразование $\ln a - \ln b = \ln a + \ln\frac{1}{b} = \ln\frac{a}{b}$, мы получим

$$\ln\frac{1+x}{1-x} = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots\right). \qquad (21)$$

Этот ряд сходится быстрее, и, кроме того, левая часть формулы может теперь выразить логарифм любого положительного числа z, так как

уравнение $\frac{1+x}{1-x} = z$ имеет при любом положительном z решение x, заключенное между -1 и $+1$. Например, если нам нужно вычислить $\ln 3$, то мы положим $x = \frac{1}{2}$ и тогда получим

$$\ln 3 = \ln \frac{1+\frac{1}{2}}{1-\frac{1}{2}} = 2\left(\frac{1}{1\cdot 2} + \frac{1}{3\cdot 2^3} + \frac{1}{5\cdot 2^5} + \ldots\right).$$

Взяв всего лишь 6 членов, вплоть до члена $\frac{2}{11\cdot 2^{11}} = \frac{1}{11264}$, мы находим значение

$$\ln 3 = 1{,}0986$$

с пятью значащими цифрами.

§ 7. Дифференциальные уравнения

1. Определения. Главенствующая роль, которую показательные и тригонометрические функции играют в математическом анализе и его применениях к задачам физики, основывается на том, что эти функции являются решениями простейших «дифференциальных уравнений».

Дифференциальным уравнением относительно неизвестной функции $u = f(x)$ с производной $u' = f'(x)$ (обозначение u' очень удачно сокращает обозначение $f'(x)$, поскольку величина u и ее формальная зависимость от x как функции $f(x)$ не нуждаются в особенном подчеркивании) называется уравнение, содержащее функцию u, производную u' и, может быть, независимое переменное x, как, например,

$$u' = u + \sin(xu)$$

или
$$u' + 3u = x^2.$$
В более общем случае дифференциальное уравнение может содержать вторую производную $u'' = f''(x)$ или производные более высокого порядка, как, например, уравнение
$$u'' + 2u' - 3u = 0.$$

Во всех подобных случаях задачей является нахождение функции $u = f(x)$, удовлетворяющей данному уравнению.

Решение дифференциальных уравнений есть широкое обобщение задачи интегрирования, понимаемой как нахождение первообразной функции по заданной функции $g(x)$: последнее сводится к решению простейшего дифференциального уравнения
$$u' = g(x).$$
Например, решениями дифференциального уравнения
$$u' = x^2$$
являются функции $u = \dfrac{x^3}{3} + c$, где c — произвольное постоянное.

2. Дифференциальное уравнение экспоненциальной функции. Радиоактивный распад. Закон роста. Сложные проценты. Показательная функция $u = e^x$ является решением дифференциального уравнения
$$u' = u, \tag{1}$$
так как производная от показательной функции равна самой показательной функции. И вообще, функция $u = ce^x$, где c — произвольное постоянное, есть решение уравнения (1). Аналогично, функция
$$u = ce^{kx}, \tag{2}$$
где c и k — две какие-нибудь постоянные, есть решение дифференциального уравнения
$$u' = ku. \tag{3}$$
Обратно, всякая функция $u = f(x)$, удовлетворяющая уравнению (3), имеет вид (2). В самом деле, пусть функции $x = h(u)$ и $u = f(x)$ взаимно обратные; в таком случае, следуя правилам дифференцирования обратной функции, найдем
$$h' = \frac{1}{u'} = \frac{1}{ku}.$$

Функцией, первообразной по отношению к найденной производной $\dfrac{1}{ku}$, является функция $\dfrac{\ln u}{k}$; итак, $x = h(u) = \dfrac{\ln u}{k} + b$, где b — некоторое постоянное. Отсюда
$$\ln u = kx - bk$$

и
$$u = e^{kx} \cdot e^{-bk}.$$
Полагая постоянную величину e^{-bk} равной c, получим
$$u = ce^{kx},$$
как и нужно было предвидеть.

Большое значение уравнения (3) заключается в том, что оно «регулирует» физические процессы, в которых количество u какого-нибудь вещества представляет собой функцию времени
$$u = f(t)$$
и притом изменяется таким образом, что скорость изменения в каждый момент пропорциональна количеству u вещества, имеющегося налицо. В этом случае *скорость изменения* в момент t, т. е.
$$u' = f'(t) = \lim \frac{f(t_1) - f(t)}{t_1 - t},$$
равна ku, где k — постоянный коэффициент пропорциональности, положительный, если u возрастает, и отрицательный, если u убывает. В обоих случаях функция u удовлетворяет дифференциальному уравнению (3); следовательно, она имеет вид
$$u = ce^{kt}.$$

Постоянная c определена, если известно количество вещества u_0, имевшееся налицо в начальный момент, т. е. при $t = 0$. Величину u_0 мы должны получить при подстановке $t = 0$ в уравнение (2):
$$u_0 = ce^0 = c;$$
отсюда и получается
$$u = u_0 e^{kt}. \tag{4}$$

Следует обратить внимание на то, что мы исходим из предположения, что задана *скорость изменения* величины u, и выводим закон (4), который позволяет вычислить фактическое количество вещества u в любой момент времени t. Эта задача как раз противоположна задаче нахождения производной от какой-нибудь функции.

Типичным примером явления указанного типа можно считать распад некоторого радиоактивного вещества. Пусть $u = f(t)$ есть количество вещества в момент времени t; если принять гипотезу, что каждая индивидуальная частица вещества имеет некоторую определенную вероятность распада и что эта вероятность не зависит от присутствия других частиц, то скорость, с которой количество u будет распадаться в данный момент

времени t, будет пропорциональна общему количеству вещества u, имеющемуся налицо. Таким образом, функция u должна удовлетворять уравнению (3) при отрицательной постоянной k, которая измеряет быстроту процесса распада; итак, вид функции следующий:

$$u = u_0 e^{kt}.$$

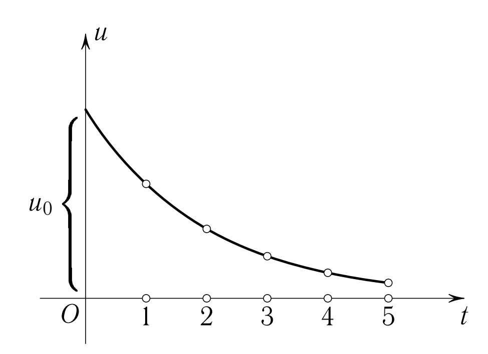


Рис. 279. Убывание по экспоненциальному закону $u = u_0 e^{kt}$, $k < 0$

Отсюда вытекает, что в равные промежутки времени подвергается распаду одна и та же доля имеющегося налицо вещества; действительно, если u_1 есть количество вещества, имеющегося в момент времени t_1, а u_2 — в некоторый последующий момент времени t_2, то

$$\frac{u_2}{u_1} = \frac{u_0 e^{kt_2}}{u_0 e^{kt_1}} = e^{k(t_2 - t_1)},$$

и последнее выражение зависит только от разности $t_2 - t_1$. Вычислим, например, сколько времени потребуется для того, чтобы в процессе распада осталась ровно половина вещества: нам нужно определить $s = t_2 - t_1$ из уравнения

$$\frac{u_2}{u_1} = \frac{1}{2} = e^{ks},$$

и мы получаем

$$ks = \ln\frac{1}{2}, \qquad s = \frac{-\ln 2}{k}. \tag{5}$$

Напротив, зная s, можно определить k:

$$k = -\frac{\ln 2}{s}.$$

Для каждого радиоактивного вещества значение s носит название «периода полураспада». Число s или некоторое аналогичное $\left(\text{например такое, как значение } r, \text{ при котором } \frac{u_2}{u_1} = \frac{999}{1000}\right)$ может быть найдено экспериментальным путем. Для радия «период полураспада» равен приблизительно 1550 годам, следовательно,

$$k = \frac{\ln\frac{1}{2}}{1550} = -0{,}0000447.$$

Отсюда мы находим, что

$$u = u_0 e^{-0{,}0000447t}.$$

Примером закона, близкого к рассмотренному только что закону показательной функции, может служить явление так называемых сложных процентов. Пусть некоторый капитал u_0 (долларов) отдан в рост из расчета 3% (сложных) в год. По истечении одного года капитал станет равным

$$u_1 = u_0(1 + 0{,}03);$$

по истечении двух лет он будет

$$u_2 = u_1(1 + 0{,}03) = u_0(1 + 0{,}03)^2;$$

наконец, по истечении t лет он выразится числом

$$u_t = u_0(1 + 0{,}03)^t. \tag{6}$$

Теперь, если начисление процентов происходило бы не один раз в год, а один раз в месяц, или, вообще, один раз в n-ю часть года, то по истечении t лет наращенный капитал выразился бы формулой

$$u_0\left(1 + \frac{0{,}03}{n}\right)^{nt} = u_0\left[\left(1 + \frac{0{,}03}{n}\right)^n\right]^t.$$

Если предположить, что число n очень велико, так что проценты присчитываются ежедневно или даже ежечасно, то, воображая, что n стремится к бесконечности, мы заметим, что величина, стоящая в скобках, стремится к $e^{0,03}$ согласно сказанному в § 6, и в пределе капитал по истечении t лет выразится формулой

$$u_0 e^{0{,}03t}, \tag{7}$$

что соответствует процессу непрерывного присчитывания сложных процентов. Можно также вычислить время s, нужное для того, чтобы удвоить основной капитал, отданный в рост по 3 сложных непрерывно начисляемых процента. Мы имеем $\dfrac{u_0 e^{0{,}03s}}{u_0} = 2$, откуда $s = \dfrac{100}{3}\ln 2 = 23{,}10$. Итак, капитал удвоился бы по истечении приблизительно 23 лет.

Вместо того чтобы шаг за шагом проделывать описанную выше процедуру и затем переходить к пределу, мы могли бы получить формулу (7), просто сказав, что скорость u' возрастания капитала u пропорциональна этому капиталу, с коэффициентом пропорциональности $k = 0{,}03$, согласно дифференциальному уравнению

$$u' = ku, \quad \text{где} \quad k = 0{,}03.$$

Тогда формула (7) вытекала бы непосредственно из общей формулы (4).

3. Другие примеры. Простые колебания. Показательная функция встречается часто в более сложных комбинациях. Например, функция

$$u = e^{-kx^2}, \tag{8}$$

где k — положительная константа, является решением дифференциального уравнения

$$u' = -2kxu.$$

Функция (8) играет большую роль в теории вероятностей и в статистике, выражая, как говорят, «нормальный» закон распределения.

Тригонометрические функции $u = \cos t$ и $v = \sin t$ также удовлетворяют простому дифференциальному уравнению. Прежде всего обратим внимание на соотношения

$$u' = -\sin t = -v,$$
$$v' = \cos t = u,$$

образующие «систему двух дифференциальных уравнений с двумя неизвестными функциями». Дифференцируя вторично, мы находим

$$u'' = -v' = -u,$$
$$v'' = u' = -v;$$

таким образом, обе функции u и v временного переменного t могут рассматриваться как решение одного и того же дифференциального уравнения

$$z'' + z = 0. \tag{9}$$

Это — очень простое дифференциальное уравнение «второго порядка», т. е. уравнение, содержащее вторую производную от функции z. Оно, или, лучше сказать, его обобщение, содержащее положительную постоянную k^2,

$$z'' + k^2 z = 0 \tag{10}$$

(решениями которого являются функции $z = \cos kt$ и $z = \sin kt$), постоянно встречается при изучении теории колебаний, и потому «синусоидальные» кривые $u = \sin kt$ и $u = \cos kt$ (рис. 280) в высшей степени интересуют всех, кто занимается конструкцией механизмов, совершающих или порождающих колебательные движения.

Следует заметить, что дифференциальное уравнение (10) представляет «идеальный» случай, когда трение или сопротивление предполагаются отсутствующими.

В дифференциальном уравнении колебательного движения сопротивление выражается лишним членом, а именно rz', так что уравнение имеет вид

$$z'' + rz' + k^2 z = 0; \tag{11}$$

его решениями являются «затухающие» колебания, выражающиеся математически с помощью формулы

$$e^{-\frac{rt}{2}}\cos\omega t \quad \text{или} \quad e^{-\frac{rt}{2}}\sin\omega t; \quad \omega = \sqrt{k^2 - \left(\frac{r}{2}\right)^2},$$

что графически представлено на рис. 281. (В качестве упражнения читатель пусть проверит правильность этих решений путем дифференцирования.) Затухающие колебания того же самого типа, что и обыкновенные синусоиды или косинусоиды, но с течением времени их размах уменьшается вследствие присутствия показательного множителя, убывающего более или менее быстро, в зависимости от величины коэффициента трения r.

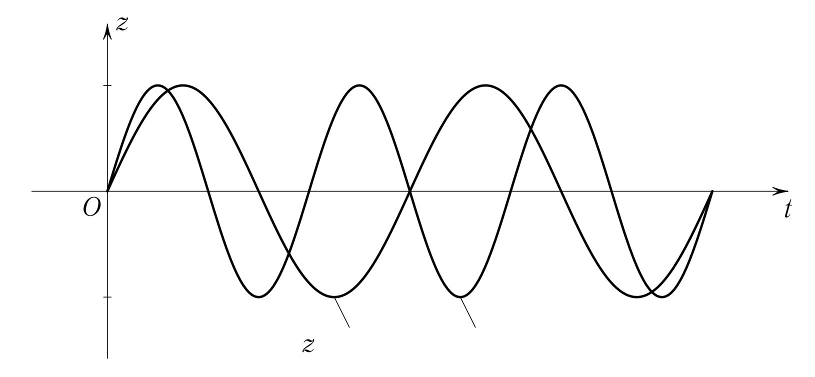


Рис. 280. Гармонические колебания

4. Закон движения Ньютона. Хотя более подробный анализ подобных явлений нами не предусмотрен, мы все же хотим включить их в общую схему, на основе которой Ньютон произвел подлинную революцию в механике и физике.

Рассмотрим вместе с Ньютоном движение некоторой частицы, имеющей массу m; обозначим ее пространственные координаты, являющиеся функциями времени t, через $x(t)$, $y(t)$, $z(t)$; таким образом, компоненты ускорения равны вторым производным $x''(t)$, $y''(t)$, $z''(t)$. В истории науки фактом решающего значения оказалось осознание Ньютоном того, что величины mx'', my'', mz'' могут быть рассматриваемы как компоненты силы, действующей на частицу. С первого взгляда может показаться, что в этой формулировке содержится всего лишь формальное определение понятия «силы». Но большой успех Ньютона заключается в том, что он первый привел это определение в соответствие с действительными явлениями природы: дело обстоит так, как будто бы сама природа предоставляла силовое «поле», которое мы можем считать известным, тогда как нам ничего не известно заранее об интересующем

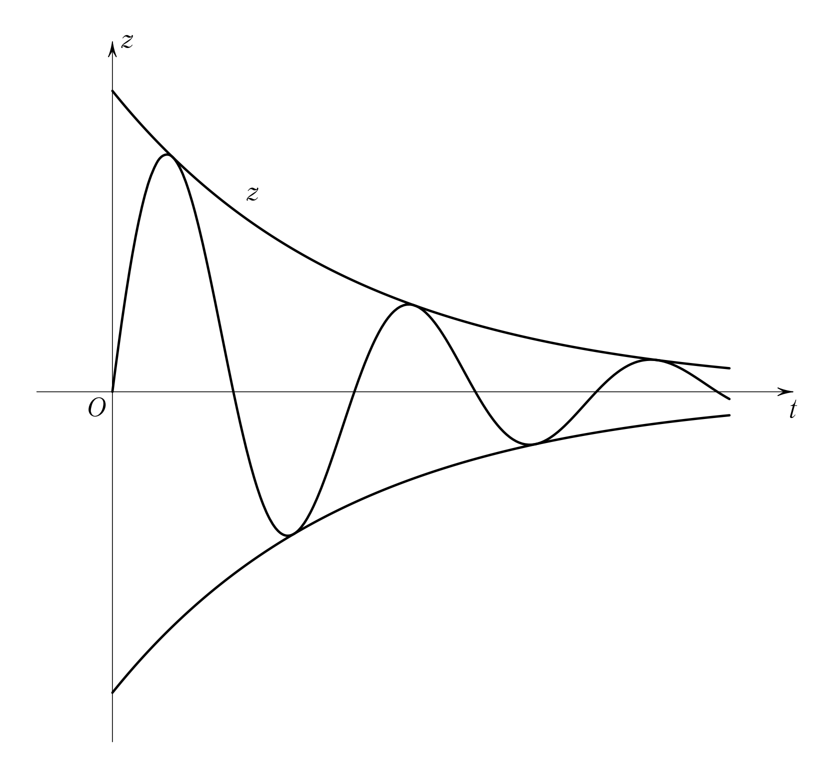


Рис. 281. Затухающее колебание

нас движении частицы в этом поле. Величайший триумф ньютоновской динамики — обоснование законов Кеплера о движении планет — ясно показывает полную гармонию между математическими концепциями Ньютона и явлениями природы. Прежде всего Ньютон предположил, что сила тяготения обратно пропорциональна квадрату расстояния. Если мы допустим, что Солнце находится в начале координатной системы, и примем величины x, y, z за координаты данной планеты, то компоненты силы по направлениям трех координатных осей будут равны соответственно

$$-k \cdot \frac{x}{r^3}, \quad -k \cdot \frac{y}{r^3}, \quad -k \cdot \frac{z}{r^3},$$

где $r = \sqrt{x^2 + y^2 + z^2}$ есть расстояние от Солнца до планеты, а k — постоянная тяготения, не зависящая от времени. Эти выражения определяют силовое поле независимо от движения в нем частицы. Известные нам данные, характеризующие это поле, нужно связать с общим ньютоновым законом движения (т. е. связать кинематические и динамические элементы); приравнивая два различных выражения вектора, мы получим

систему трех дифференциальных уравнений

$$mx'' = \frac{-kx}{(x^2+y^2+z^2)^{3/2}},$$
$$my'' = \frac{-ky}{(x^2+y^2+z^2)^{3/2}},$$
$$mz'' = \frac{-kz}{(x^2+y^2+z^2)^{3/2}}$$

с тремя неизвестными функциями $x(t)$, $y(t)$, $z(t)$. Эту систему можно решить, и тогда обнаружится, что в полном согласии с кеплеровскими эмпирическими наблюдениями орбита планеты есть коническое сечение с Солнцем в одном из фокусов, что площади, описываемые в равные промежутки времени радиусом-вектором, проведенным от Солнца к планете, равны между собой и что квадраты периодов полного обращения двух планет вокруг Солнца пропорциональны кубам их расстояний от Солнца. Доказательство этих утверждений мы вынуждены опустить.

Задача о колебательном движении предоставляет более элементарную иллюстрацию метода Ньютона. Предположим, что мы имеем частицу, движущуюся по прямой линии, по оси x, и связанную с началом координат силой упругости, что можно, например, осуществить с помощью пружины или резинки.

Если частица выведена из положения равновесия (в начале координат) и помещена в некоторую точку с координатой x, то сила потянет ее назад. Мы предположим, что эта сила пропорциональна растяжению x; так как она направлена к началу координат, то представится в виде $-k^2x$, где $-k^2$ — отрицательный множитель пропорциональности, выражающий силу упругости пружины или резинки.

Мы предположим, далее, что налицо имеется трение, замедляющее движение, и что это трение пропорционально скорости x' частицы с коэффициентом пропорциональности, равным $-r$. Тогда результирующая сила в любой момент времени выразится через $-k^2x - rx'$, и, пользуясь общим принципом Ньютона, мы приходим к уравнению $mx'' = -k^2x - rx'$, или

$$mx'' + rx' + k^2x = 0.$$

А это — не что иное, как рассмотренное выше дифференциальное уравнение (11) затухающих колебаний. Предыдущий простой пример имеет большое значение, так как многие колебания механических или электрических систем могут быть математически записаны с помощью именно этого дифференциального уравнения. Здесь мы имеем типичный пример того, как отвлеченная математическая формулировка одним ударом обнажает внутреннюю структуру многих, казалось бы, совершенно различных и не связанных между собой отдельных явлений. Подобного рода абстрагирование от частного характера данного явления и пере-

ход к общему закону, регулирующему обширный класс явлений, есть характерная черта математической трактовки физических проблем.

ДОПОЛНЕНИЕ К ГЛАВЕ VIII

§ 1. Вопросы принципиального порядка

1. Дифференцируемость. Понятие производной от некоторой функции $y = f(x)$ мы связывали с интуитивным представлением о касательной к графику этой функции. Но так как общая концепция функции чрезвычайно широка, то в интересах логической законченности необходимо уничтожить эту зависимость от геометрической интуиции. В самом деле, мы ведь не гарантированы от того, что интуитивные свойства, бросающиеся в глаза при рассмотрении простых кривых, подобных кругу или эллипсу, не исчезнут для графиков более сложных функций. Рассмотрим, например, функцию, изображенную на рис. 282, график которой имеет угловую точку.

Эта функция определяется уравнением $y = x + |x|$, где символом $|x|$ обозначается абсолютная величина x; иными словами,

$$y = x + x = 2x \qquad \text{при } x \geqslant 0,$$
$$y = x - x = 0 \qquad \text{при } x < 0.$$

Другим примером такого рода может служить функция $y = |x|$, а также функция $y = x + |x| + (x - 1) + |x - 1|$. Графики этих функций в некоторых точках перестают иметь определенную касательную, т. е. определенное направление; это значит, что функция в соответствующих точках x не имеет производной.

Упражнения. 1) Постройте (т. е. запишите с помощью конкретных аналитических выражений) функцию $f(x)$, график которой есть половина правильного шестиугольника.

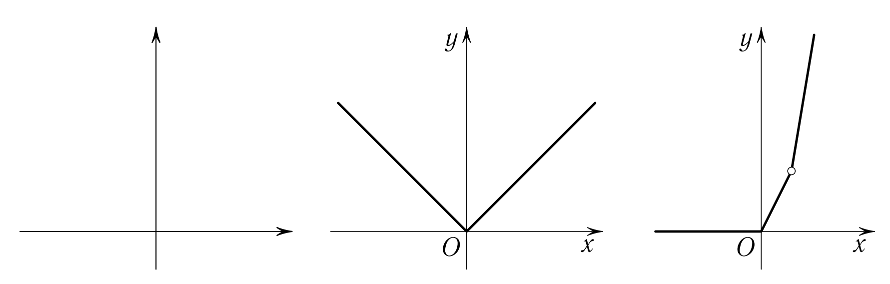


Рис. 282. $y = x + |x|$

Рис. 283. $y = |x|$

Рис. 284. $y = x + |x| + (x - 1) + |x - 1|$

2) Где расположены угловые точки графика

$$f(x) = (x + |x|) + \frac{1}{2}\left\{\left(x - \frac{1}{2}\right) + \left|x - \frac{1}{2}\right|\right\} + \frac{1}{4}\left\{\left(x - \frac{1}{4}\right) + \left|x - \frac{1}{4}\right|\right\}?$$

Каковы точки разрыва производной $f'(x)$?

В качестве простого примера недифференцируемости уже иного типа приведем функцию

$$y = f(x) = x \sin\frac{1}{x},$$

получаемую посредством умножения функции $\sin\frac{1}{x}$ (см. стр. 304) на множитель x; положим, по определению, что $f(x) = 0$ при $x = 0$. Эта функция, график которой для положительных значений переменного x изображен на рис. 285, непрерывна в каждой точке. График колеблется бесконечно часто в окрестности точки $x = 0$, причем «волны» становятся бесконечно малыми, если мы приближаемся к нулю. Наклон этих волн дается формулой

$$f'(x) = \sin\frac{1}{x} - \frac{1}{x}\cos\frac{1}{x}$$

(пусть читатель проверит это в качестве упражнения); при стремлении x к нулю этот наклон колеблется между все возрастающими положительной и отрицательной границами. Мы можем сделать попытку найти производную в точке $x = 0$, переходя к пределу при $h \to 0$ в разностном отношении

$$\frac{f(0+h) - f(0)}{h} = \frac{h \sin\frac{1}{h}}{h} = \sin\frac{1}{h}.$$

Но при $h \to 0$ это разностное отношение колеблется между -1 и $+1$ и не стремится ни к какому пределу, следовательно, функция не может быть продифференцирована в точке $x = 0$.

Эти примеры указывают на трудности, внутренне присущие самому вопросу. Вейерштрасс удивительно ярко проиллюстрировал положение вещей, построив непрерывную функцию, график которой не имеет производной ни в одной точке. В то время как дифференцируемость влечет за собой непрерывность, непрерывность, как показывает этот пример, отнюдь не влечет за собой дифференцируемости; в самом деле, функция Вейерштрасса всюду непрерывна, а вместе с тем нигде не дифференцируема. На практике трудности такого рода не встретятся. Обычно встречающиеся кривые являются «гладкими» (за исключением разве только отдельных изолированных точек), т. е. дифференцирование не только возможно, но даже сама производная является непрерывной. Что же в таком случае может нам помешать просто сделать оговорку, что никакие «патологические» явления не будут фигурировать в задачах, подлежащих нашему рассмотрению? Именно так и поступают в анализе те, кому приходится иметь дело только с

дифференцируемыми функциями. В главе VIII мы провели дифференцирование обширного класса функций и тем самым доказали их дифференцируемость.

Поскольку дифференцируемость функции не есть логическая неизбежность, она — с математической точки зрения — должна быть или постулирована, или доказана. В таком случае само понятие касательной или направления кривой (первоначальный источник идеи производной) ставится в зависимость от чисто аналитического определения производной: если функция $y = f(x)$ дифференцируема, т. е. если разностное отношение $\dfrac{f(x+h) - f(x)}{h}$ имеет единственный предел $f'(x)$ при стремлении h к нулю с обеих сторон, то принято говорить, что соответствующая кривая имеет касательную с наклоном $f'(x)$. Таким образом, наивная позиция Ферма, Лейбница и Ньютона в современном анализе вывернута наоборот — в интересах логической стройности.

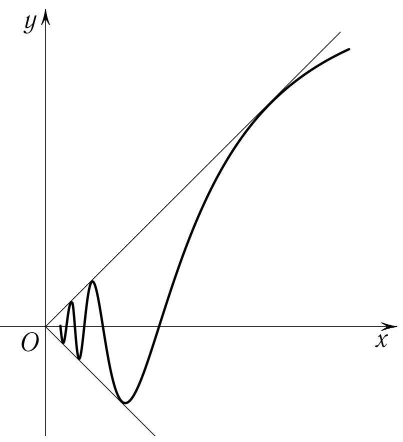


Рис. 285. $y = x \sin \dfrac{1}{x}$

Упражнения. 1) Покажите, что непрерывная функция, определенная формулой $f(x) = x^2 \sin \dfrac{1}{x}$ и добавочным условием $f(0) = 0$, дифференцируема в точке $x = 0$.

2) Покажите, что функция $\operatorname{arctg} \dfrac{1}{x}$ разрывна при $x = 0$, что функция $x \operatorname{arctg} \dfrac{1}{x}$ непрерывна в этой точке, но не имеет в ней производной, и что функция $x^2 \operatorname{arctg} \dfrac{1}{x}$ дифференцируема при $x = 0$. (В двух последних примерах следует положить значение функций при $x = 0$ равным нулю.)

2. Интеграл. Аналогично и положение с интегралом от непрерывной функции $f(x)$. Вместо того чтобы «площадь под кривой» принимать как величину, объективно существующую и которую a posteriori можно выразить с помощью предела последовательности конечных сумм, этот предел в анализе принимают в качестве *определения* интеграла. Эта концепция интеграла образует первичную основу, из которой затем выводится понятие площади. Мы вынуждены стать на эту точку зрения вследствие сознания того, что геометрическая интуиция обладает известной расплывчатостью, когда она применяется к таким общим аналитическим понятиям, как непрерывная функция. Мы начнем с по-

строения суммы

$$S_n = \sum_{j=1}^{n} f(v_j)(x_j - x_{j-1}) = \sum_{j=1}^{n} f(v_j)\Delta x_j, \tag{1}$$

где $x_0 = a, x_1, \ldots, x_n = b$ — точки деления промежутка интегрирования, $\Delta x_j = x_j - x_{j-1}$ — приращение переменной x, или длина j-го частного промежутка, а v_j — произвольное значение переменного x в этом частном промежутке, т. е. $x_{j-1} \leqslant v_j \leqslant x_j$ (мы можем взять, например, $v_j = x_j$ или $v_j = x_{j-1}$.)

Далее, мы образуем последовательность подобных сумм, в которых число n частных промежутков возрастает, причем длина максимального частного промежутка стремится к нулю. Тогда справедливо следующее основное положение: сумма S_n, составленная для данной непрерывной функции $f(x)$, стремится к некоторому определенному пределу A, не зависящему от способа разбиения промежутка интегрирования и от выбора точек v_j. По определению, этот предел есть интеграл $A = \int\limits_a^b f(x)\,dx$. Конечно, существование этого предела должно быть аналитически доказано, если мы не хотим ссылаться на интуитивное геометрическое представление площади. Это доказательство приводится в каждом руководстве анализа, учитывающем требования математической строгости.

Сравнение дифференцирования и интегрирования приводит нас к следующему противопоставлению. Свойство дифференцируемости, несомненно, налагает ограничительное условие на класс всех непрерывных функций; вместе с тем фактическое выполнение операции дифференцирования сводится на практике к процедурам, основанным лишь на нескольких простых правилах. В противоположность этому каждая непрерывная функция без исключения интегрируема, так как обладает интегралом между любыми двумя данными пределами. Однако прямое вычисление интегралов, понимаемых как пределы сумм, даже в случае самых простых функций, вообще говоря, дело очень трудное. Но тут-то и оказывается, что основная теорема анализа во многих случаях становится решающим орудием при осуществлении интегрирования. И все же для большей части функций, в том числе даже для некоторых совершенно элементарных, интегрирование не дает простых явных выражений, и числовые выкладки для интегралов требуют более продвинутых методов.

3. Другие приложения понятия интеграла. Работа. Длина кривой. Оторвав аналитическое представление интеграла от его первоначальной геометрической интерпретации, мы встречаемся с целым рядом других, не менее важных интерпретаций и приложений этого основного понятия. Например, в механике интеграл может быть интер-

претирован как выражение работы. Достаточно будет разъяснить это на следующем простом примере. Предположим, что некоторая масса движется по оси x под влиянием силы, направленной вдоль этой оси. Будем считать, что вся масса сосредоточена в одной точке с координатой x и что сила задана как функция этой точки $f(x)$, причем знак функции $f(x)$ указывает на направление силы. Если сила постоянна и передвигает массу из точки a в точку b, то работа, произведенная ею, равна произведению величины силы f на пройденный массой путь: $(b-a)f$. Но если сила меняется вместе с изменением x, то придется определять общую произведенную работу с помощью предельного процесса (подобно тому, как мы прежде определяли скорость). Для этой цели мы разобьем промежуток от a до b, как и прежде, на мелкие частные промежутки точками $x_0=a$, x_1, x_2, ..., $x_n=b$; затем предположим, что в каждом частном промежутке сила остается постоянной и равной, скажем, величине $f(x_\nu)$, истинному значению силы в конечной точке, и вычислим работу, соответствующую такой «ступенчатой» силе:

$$S_n=\sum_{\nu=1}^{n} f(x_\nu)\Delta x_\nu.$$

Если мы теперь, как раньше, станем уменьшать промежутки деления, заставляя n неограниченно расти, мы увидим, что сумма будет стремиться к интегралу

$$\int_a^b f(x)\,dx.$$

Таким образом, работа, совершаемая непрерывно меняющейся силой, определена с помощью интеграла.

В частности, рассмотрим массу m, связанную с началом координат $x=0$ упругой пружиной. Сила $f(x)$, согласно рассуждению на стр. 486, будет пропорциональна x,

$$f(x)=-k^2x,$$

где k^2 — положительная постоянная. Тогда работа, совершенная этой силой при перемещении массы m из начала координат в точку b, выразится интегралом

$$\int_a^b (-k^2x)\,dx=-k^2\,\frac{b^2}{2},$$

а работа, которую мы сами должны затратить при растяжении пружины до точки b, равна $+k^2\,\dfrac{b^2}{2}$.

Другое приложение общего понятия интеграла — это вычисление длины дуги кривой. Предположим, что рассматриваемая часть кривой

представлена функцией $y = f(x)$, производная которой $f'(x) = \dfrac{dy}{dx}$ также непрерывная функция. Для того чтобы определить длину, мы будем действовать точно так, как если бы нам надо было измерить длину кривой для практических целей при помощи линейки с делениями. Впишем в дугу AB ломаную линию с n маленькими сторонами, измерим общую длину (периметр) L_n этой ломаной и станем рассматривать эту длину как некоторое приближение; заставим n возрастать, а наибольшую из сторон ломаной — стремиться к нулю; тогда мы получим в качестве длины дуги AB следующий предел:

$$L = \lim L_n.$$

(В главе VI этим же способом была получена длина окружности как предел периметра вписанного правильного n-угольника.) Можно доказать, что для достаточно гладких кривых этот предел существует и не зависит от того, каким образом выбирается последовательность вписанных ломаных. Те кривые, для которых это имеет место, называются *спрямляемыми*. Всякая «порядочная» кривая, встречающаяся в теории или ее приложениях, оказывается спрямляемой, и мы не станем углубляться в исследование «патологических» случаев. Достаточно будет показать, что дуга AB для функции $y = f(x)$ с непрерывной производной $f'(x)$ имеет длину L в указанном смысле и что длина L может быть выражена с помощью интеграла.

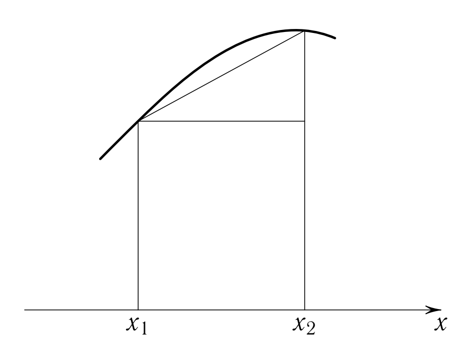


Рис. 286. К определению длины дуги

С этой целью обозначим абсциссы точек A и B соответственно через a и b, затем разобьем промежуток от a до b, как и прежде, точками $a = x_0, x_1, x_2, \ldots, x_n = b$ с разностями $\Delta x_j = x_j - x_{j-1}$ и рассмотрим ломаную линию с вершинами $(x_j,\ y_j = f(x_j))$, расположенными над точками деления. Длина одной из сторон ломаной выразится формулой

$$\sqrt{(x_j - x_{j-1})^2 + (y_j - y_{j-1})^2} = \sqrt{\Delta x_j^2 + \Delta y_j^2} = \Delta x_j \cdot \sqrt{1 + \left(\frac{\Delta y_j}{\Delta x_j}\right)^2}.$$

Отсюда для общей длины ломаной линии получается выражение

$$L_n = \sum_{j=1}^{n} \sqrt{1 + \left(\frac{\Delta y_j}{\Delta x_j}\right)^2}\,\Delta x_j.$$

Если заставить теперь n стремиться к бесконечности, то разностное от-

ношение $\dfrac{\Delta y_j}{\Delta x_j}$ будет стремиться к производной $\dfrac{dy}{dx} = f'(x)$, и мы получим для длины L интегральное выражение

$$L = \int\limits_a^b \sqrt{1 + [f'(x)]^2}\, dx. \qquad (2)$$

Не вдаваясь в дальнейшие подробности этих теоретических рассуждений, мы сделаем два дополнительных замечания. Во-первых, если точку B считать подвижной точкой на данной кривой с абсциссой x, то $L = L(x)$ становится функцией переменного x, и, согласно основной теореме, мы имеем формулу

$$L'(x) = \frac{dL}{dx} = \sqrt{1 + [f'(x)]^2},$$

которую часто приходится применять. Во-вторых, хотя формула (2) и дает «общее» решение задачи нахождения длины дуги, все же она редко позволяет найти явное выражение этой длины в отдельных частных случаях. В самом деле, чтобы получить числовое значение длины дуги, мы должны подставить данную функцию $f(x)$, или, точнее, $f'(x)$, в формулу (2) и тогда осуществить фактическое интегрирование полученного выражения. Но здесь возникают, вообще говоря, непреодолимые трудности, если мы ограничим себя областью элементарных функций, рассмотренных в этой книге. Укажем небольшое число случаев, для которых интегрирование возможно. Функция

$$y = f(x) = \sqrt{1 - x^2}$$

имеет графиком единичный круг; для нее мы получаем

$$f'(x) = \frac{dy}{dx} = -\frac{x}{\sqrt{1-x^2}}, \quad \text{откуда} \quad \sqrt{1 + [f'(x)]^2} = \frac{1}{\sqrt{1-x^2}};$$

следовательно, длина дуги окружности выражается интегралом

$$\int\limits_a^b \frac{dx}{\sqrt{1-x^2}} = \arcsin b - \arcsin a.$$

Для случая параболы $y = x^2$ мы имеем $f'(x) = 2x$, а длина дуги от $x = 0$ до $x = b$ равна

$$\int\limits_a^b \sqrt{1 + 4x^2}\, dx.$$

Для кривой $y = \ln\sin x$ мы имеем $f'(x) = \operatorname{ctg} x$, и длина дуги выражается интегралом

$$\int\limits_a^b \sqrt{1 + \operatorname{ctg}^2 x}\, dx.$$

Мы удовольствуемся лишь простым написанием этих интегральных выражений. Их можно было бы вычислить, применяя несколько более развитую технику интегрирования, чем та, которая имеется в нашем распоряжении, но мы не пойдем дальше в этом направлении.

§ 2. Порядки возрастания

1. Показательная функция и степени переменного x. В математике мы постоянно встречаемся с последовательностями чисел a_n, которые имеют бесконечный предел. Часто бывает нужно сравнить такую последовательность с другой последовательностью, например, чисел b_n, тоже стремящихся к бесконечности, но, может быть, «быстрее», чем последовательность чисел a_n. Уточним это понятие: мы скажем, что b_n стремится к бесконечности быстрее, чем a_n, или b_n имеет *более высокий порядок возрастания*, чем a_n, если *отношение* $\frac{a_n}{b_n}$ (в котором как числитель, так и знаменатель стремятся к бесконечности) стремится к нулю при возрастании n. Например, последовательность $b_n = n^2$ стремится к бесконечности быстрее, чем последовательность $a_n = n$, а эта последовательность, в свою очередь, быстрее, чем последовательность $c_n = \sqrt{n}$, так как

$$\frac{a_n}{b_n} = \frac{n}{n^2} = \frac{1}{n} \to 0, \qquad \frac{c_n}{a_n} = \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}} \to 0.$$

Ясно, что n^s стремится к бесконечности быстрее чем n^r (при $s > r > 0$), так как $\frac{n^r}{n^s} = \frac{1}{n^{s-r}} \to 0$.

Если отношение $\frac{a_n}{b_n}$ стремится к некоторой конечной постоянной c, отличной от нуля, то мы говорим, что обе последовательности a_n и b_n стремятся к бесконечности с *одинаковой скоростью* или что они имеют *одинаковый порядок возрастания.* Так, например, $a_n = n^2$ и $b_n = 2n^2 + n$ имеют один и тот же порядок возрастания, потому что

$$\frac{a_n}{b_n} = \frac{n^2}{2n^2 + n} = \frac{1}{2 + \frac{1}{n}} \to \frac{1}{2}.$$

Могла бы возникнуть мысль, что возрастание любой последовательности a_n с бесконечным пределом может быть «измерено» с помощью степеней n^s так же, как любой отрезок может быть измерен с помощью линейки с делениями. Стоило бы только для этого найти подходящую степень n^s с тем же порядком возрастания, что и a_n, т. е. такую, что отношение $\frac{a_n}{n^s}$ стремится к некоторой конечной, отличной от нуля постоянной. Но совершенно замечательным является то обстоятельство, что осуществить это отнюдь не всегда возможно — хотя бы потому, что *показательная функция a^n при $a > 1$ (например, e^n) стремится к бесконечности быстрее, чем какая бы то ни было степень n^s, как бы велик*

ни был показатель s; *с другой стороны, функция* $\ln n$ *стремится к бесконечности медленнее, чем какая бы то ни было степень* n^s, *как бы мал ни был положительный показатель* s. Другими словами, мы имеем соотношения

$$\frac{n^s}{a^n} \to 0 \tag{1}$$

и

$$\frac{\ln n}{n^s} \to 0 \tag{2}$$

при $n \to \infty$. Заметим, что показатель степени s — не обязательно целое число; он может быть любым фиксированным *положительным* числом.

Для того чтобы доказать соотношение (1), мы упростим наше утверждение тем, что извлечем из соотношения (1) корень степени s; ясно, что вместе с корнем стремится к нулю и подкоренное выражение. Итак, нам остается только доказать, что

$$\frac{n}{a^{\frac{n}{s}}} \to 0$$

при возрастании n. Пусть $b = a^{\frac{1}{s}}$; так как по предположению a больше единицы, то и b и $\sqrt{b} = b^{\frac{1}{2}}$ также больше 1. Можно написать

$$b^{\frac{1}{2}} = 1 + q,$$

где q положительно. Теперь, в силу неравенства (6) на стр. 34,

$$b^{\frac{n}{2}} = (1+q)^n \geqslant 1 + nq > nq,$$

так что

$$a^{\frac{n}{s}} = b^n > n^2q^2,$$

и следовательно,

$$\frac{n}{a^{\frac{n}{s}}} < \frac{n}{n^2q^2} = \frac{1}{nq^2}.$$

Так как выражение справа стремится к нулю при $n \to \infty$, доказательство закончено.

Нужно заметить, что соотношение

$$\frac{x^s}{a^x} \to 0 \tag{3}$$

остается в силе, когда x стремится к бесконечности *любым способом*, пробегая последовательность $x_1, x_2, \ldots$, которая может и не совпадать с последовательностью 1, 2, 3, ... целых положительных чисел. В самом деле, при $n - 1 \leqslant x \leqslant n$ мы имеем

$$\frac{x^s}{a^x} < \frac{n^s}{a^{n-1}} = a \cdot \frac{n^s}{a^n} \to 0.$$

Это замечание можно использовать для доказательства соотношения (3). Если положить $x = \ln n$ и $e^s = a$, так что $n^s = (e^s)^x$, то дробь в левой части (2) примет вид

$$\frac{x}{a^x},$$

и мы приходим к выражению, имеющемуся в соотношении (3) при $s = 1$.

Упражнения. 1) Докажите, что при $x \to \infty$ функция $\ln \ln x$ стремится к бесконечности медленней, чем $\ln x$.

2) Производная от функции $\frac{x}{\ln x}$ равна разности $\frac{1}{\ln x} - \frac{1}{(\ln x)^2}$. Докажите, что при возрастании x производная «асимптотически равна» первому члену, $\frac{1}{\ln x}$, т. е. что отношение упомянутых величин при $x \to \infty$ стремится к 1.

2. Порядок возрастания функции $\ln(n!)$. Во многих приложениях, например в теории вероятностей, важно знать порядок возрастания или «асимптотическое поведение» выражения $n!$ при очень больших значениях n. Займемся здесь изучением логарифма от $n!$, т. е. выражения

$$P_n = \ln 2 + \ln 3 + \ldots + \ln n.$$

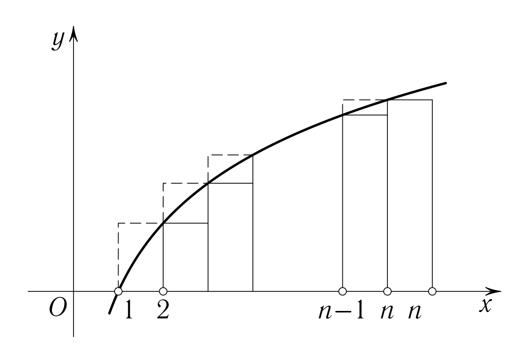


Рис. 287. Оценка $\ln(n!)$

Мы покажем, что в качестве «асимптотического значения» выражения P_n может служить произведение $n \ln n$, т. е. что

$$\frac{\ln(n!)}{n \ln n} \to 1$$

при $n \to \infty$.

Проведем доказательство так, как это обыкновенно делается, когда нужно сравнить сумму с интегралом. На рис. 287 сумма P_n равна сумме площадей прямоугольников, верхние стороны которых обозначены сплошными линиями и общая площадь которых не превосходит площади

$$\int\limits_0^{n+1} \ln x \, dx = (n+1)\ln(n+1) - (n+1) + 1$$

под логарифмической кривой в пределах от 1 до $n + 1$ [см. стр. 532, упражнение а)]. Но в то же самое время сумма P_n равна сумме площадей прямоугольников, верхние стороны которых обозначены пунктиром и общая площадь которых превосходит площадь под той же кривой в пределах от 1 до n,

$$\int\limits_1^{n} \ln x \, dx = n \ln n - n + 1.$$

Отсюда мы имеем

$$n \ln n - n + 1 < P_n < (n+1)\ln(n+1) - n;$$

разделив это неравенство на $n \ln n$, получим

$$1 - \frac{1}{\ln n} + \frac{1}{n \ln n} < \frac{P_n}{n \ln n} < \left(1 + \frac{1}{n}\right) \frac{\ln(n+1)}{\ln n} - \frac{1}{\ln n} =$$

$$= \left(1 + \frac{1}{n}\right) \frac{\ln n + \ln\left(1 + \frac{1}{n}\right)}{\ln n} - \frac{1}{\ln n}.$$

Очевидно, и верхняя и нижняя границы, между которыми заключено отношение $\frac{P_n}{n \ln n}$, стремятся к единице, и таким образом наше утверждение доказано.

Упражнение. Докажите, что упомянутые выше границы, соответственно, больше чем $1 - \frac{1}{n}$ и меньше чем $1 + \frac{1}{n}$.

§ 3. Бесконечные ряды и бесконечные произведения

1. Бесконечные ряды функций. Мы не раз уже имели случай указать, что, выражая величину s в виде «суммы бесконечного ряда»

$$s = b_1 + b_2 + b_3 + \ldots, \tag{1}$$

мы не утверждаем ничего иного, кроме того, что s есть предел при возрастающем n последовательности конечных «частных сумм»

$$s_1, \quad s_2, \quad s_3, \quad \ldots,$$

где

$$s_n = b_1 + b_2 + b_3 + \ldots + b_n. \tag{2}$$

Таким образом, равенство (1) равносильно предельному соотношению

$$\lim s_n = s \quad \text{при} \quad n \to \infty, \tag{3}$$

где s_n определено с помощью (2). Если предел (3) существует, то мы говорим, что ряд (1) *сходится* к значению s; напротив, если предел (3) не существует, то мы говорим, что этот ряд *расходится*.

Например, ряд

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$$

сходится к значению $\frac{\pi}{4}$, а ряд

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots$$

сходится к значению $\ln 2$; но, напротив, ряд

$$1 - 1 + 1 - 1 + \ldots$$

расходится, так как частные суммы здесь равны поочередно то 1, то 0; а ряд

$$1 + 1 + 1 + 1 + \ldots$$

расходится по той причине, что частные суммы стремятся к бесконечности.

Нам приходилось уже встречаться с рядами, общий член которых есть функция переменной x, имеющая вид

$$b_i = c_i x^i,$$

причем c_i не зависит от x. Такие ряды называются *степенными*; для них частными суммами являются многочлены

$$s_n = c_0 + c_1 x + c_2 x^2 + \ldots + c_n x^n;$$

прибавление постоянного члена c_0 потребует лишь несущественного изменения обозначений в формуле (2).

Разложение функции $f(x)$ в степенной ряд

$$f(x) = c_0 + c_1 x + c_2 x^2 + \ldots$$

есть, таким образом, один из способов представить функцию $f(x)$ приближенно с помощью простейших функций — полиномов. Подводя итоги предыдущим результатам и несколько дополняя их, составим следующий список уже известных нам разложений в степенные ряды:

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \ldots \qquad \text{(справедливо при } -1 < x < +1\text{)}, \tag{4}$$

$$\operatorname{arctg} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \ldots \qquad \text{(справедливо при } -1 \leqslant x \leqslant +1\text{)}, \tag{5}$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots \qquad \text{(справедливо при } -1 < x \leqslant +1\text{)}, \tag{6}$$

$$\frac{1}{2}\ln\frac{1+x}{1-x} = x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots \qquad \text{(справедливо при } -1 < x < +1\text{)}. \tag{7}$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} \ldots \qquad \text{(справедливо при всех } x\text{)}. \tag{8}$$

Сюда же мы присоединим еще два важных разложения

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots \qquad \text{(справедливо при всех } x\text{)}, \tag{9}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots \qquad \text{(справедливо при всех } x\text{)}. \tag{10}$$

Доказательство этих разложений может быть построено как простое следствие из соотношений (см. стр. 464)

$$\text{а)} \int\limits_0^x \sin u\,du = 1 - \cos x, \qquad \text{б)} \int\limits_0^x \cos u\,du = \sin x.$$

Мы отправляемся от следующего очевидного неравенства:

$$\cos x \leqslant 1.$$

Интегрируя от 0 до x, где x есть некоторое фиксированное положительное число, мы находим по формуле (13) со стр. 437:

$$\sin x \leqslant x;$$

интегрируя это еще раз, получим

$$1 - \cos x \leqslant \frac{x^2}{2},$$

что равносильно

$$\cos x \geqslant 1 - \frac{x^2}{2}.$$

Проинтегрировав последнее неравенство, найдем

$$\sin x \geqslant x - \frac{x^3}{2 \cdot 3} = x - \frac{x^3}{3!}.$$

Продолжая таким же способом до бесконечности, мы получаем две серии неравенств:

$$\begin{aligned}
&\sin x \leqslant x, && \cos x \leqslant 1,\\
&\sin x \geqslant x - \frac{x^3}{3!}, && \cos x \geqslant 1 - \frac{x^2}{2!},\\
&\sin x \leqslant x - \frac{x^3}{3!} + \frac{x^5}{5!}, && \cos x \leqslant 1 - \frac{x^2}{2!} + \frac{x^4}{4!},\\
&\sin x \geqslant x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}, && \cos x \geqslant 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!},\\
&\dots\dots\dots\dots && \dots\dots\dots\dots
\end{aligned}$$

Установим теперь, что при неограниченном возрастании n имеет место соотношение $\dfrac{x^n}{n!} \to 0$.

Для того чтобы это доказать, выберем некоторое фиксированное число m, такое что $\dfrac{x}{m} < \dfrac{1}{2}$, и введем обозначение $c = \dfrac{x^m}{m!}$. Любое целое $n > m$ представим в виде суммы $n = m + r$, тогда

$$0 < \frac{x^n}{n!} = c \cdot \frac{x}{m+1} \cdot \frac{x}{m+2} \cdot \ldots \cdot \frac{x}{m+r} < c \cdot \left(\frac{1}{2}\right)^r;$$

так как из того, что $n \to \infty$, следует, что $r \to \infty$, то $c \cdot \left(\dfrac{1}{2}\right)^r \to 0$. Отсюда и вытекает, что справедливы тождества

$$\begin{cases} \sin x = x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \dfrac{x^7}{7!} + \ldots, \\[2ex] \cos x = 1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} - \dfrac{x^6}{6!} + \ldots \end{cases}$$

Поскольку члены этих рядов, меняя поочередно знаки, убывают по величине (по крайней мере, при $|x| \leqslant 1$), то *ошибки, совершаемые при обрывании каждого из рядов на некотором члене, не превышают по абсолютной величине первого из отброшенных членов.*

Эти ряды можно использовать при составлении таблиц.

Пример. Чему равен $\sin 1°$? $1°$ равен в радианном измерении числу $\frac{\pi}{180}$; следовательно,

$$\sin\frac{\pi}{180} = \frac{\pi}{180} - \frac{1}{6}\left(\frac{\pi}{180}\right)^3 + \ldots$$

Если ограничиться выписанными двумя членами, то совершаемая при этом ошибка не будет превышать числа $\frac{1}{120}\left(\frac{\pi}{180}\right)^5$, которое меньше чем 0,00000000002. Итак, $\sin 1° \approx 0{,}0174524064$, с 10 десятичными знаками.

Наконец, упомянем без доказательства о «биномиальном ряде»

$$(1+x)^a = 1 + ax + C_2^a x^2 + C_3^a x^3 + \ldots, \tag{11}$$

где C_s^a — «биномиальный коэффициент»

$$C_s^a = \frac{a(a-1)(a-2)\ldots(a-s+1)}{s!}.$$

Если $a = n$ есть целое положительное число, то $C_n^a = 1$, и в формуле (11) все коэффициенты C_s^a при $s > n$ обращаются в нуль, так что мы просто получаем конечную формулу обыкновенной биномиальной теоремы. Одно из крупных открытий Ньютона, сделанных им в начале его деятельности, заключалось в том, что он обобщил биномиальную теорему на случай всех возможных значений показателя a как положительных, так и отрицательных, как рациональных, так и иррациональных. Если a не есть целое положительное число, то правая часть формулы (11) дает бесконечный ряд, сходящийся к значению, равному левой части, при $-1 < x < +1$. Если же $|x| > 1$, то ряд (11) расходится, и знак равенства теряет всякий смысл.

В частности, подставляя в формулу (11) значение $a = \frac{1}{2}$, мы найдем разложение

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{2!\,2^2}x^2 + \frac{1\cdot 3}{3!\,2^3}x^3 - \frac{1\cdot 3\cdot 5}{4!\,2^4}x^4 + \ldots \tag{12}$$

Подобно другим математикам XVIII в., Ньютон не дал настоящего доказательства своей формулы. Удовлетворительный анализ сходимости и пределы, в которых разложение оказывается справедливым, не были установлены для подобных рядов вплоть до XIX в.

Упражнение. Напишите степенные ряды, в которые разлагаются функции $\sqrt{1-x^2}$ и $\frac{1}{\sqrt{1-x}}$.

Разложения (4)–(11) являются частными случаями общей формулы Брука Тейлора (1685–1731), дающей разложение функции $f(x)$ в степенной ряд вида

$$f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots \tag{13}$$

Подмечая закон, выражающий коэффициенты этого ряда c_j с помощью функции $f(x)$ и ее производных, можно утверждать справедливость этого разложения для очень обширного класса функций.

Здесь невозможно привести строгое доказательство формулы Тейлора; невозможно также точно сформулировать условия, при которых она справедлива. Но следующие общедоступные соображения прольют некоторый свет на относящиеся сюда взаимоотношения и существенные факты.

Допустим предварительно, что разложение (13) возможно. Далее предположим, что функция $f(x)$ дифференцируема, что ее производная $f'(x)$ дифференцируема, и так далее, так что существует бесконечная последовательность производных

$$f'(x),\ \ f''(x),\ \ \ldots,\ \ f^{(n)}(x),\ \ \ldots$$

Наконец, будем дифференцировать бесконечный степенной ряд почленно точно так, как если бы это был конечный многочлен, не озабочиваясь вопросом о законности такой процедуры. После всех этих допущений можно определить коэффициенты c_n, зная поведение функции $f(x)$ в окрестности точки $x=0$. Прежде всего, подставляя в формулу (13) $x=0$, мы находим

$$c_0 = f(0),$$

так как все члены ряда, содержащие переменное x, исчезают.

Дифференцируя тождество (13), мы получаем

$$f'(x) = c_1 + 2c_2x + 3c_3x^2 + \ldots + nc_nx^{n-1} + \ldots; \qquad (13')$$

снова подставляя значение $x=0$, но на этот раз в формулу (13′), мы находим

$$c_1 = f'(0).$$

Дифференцируя (13′), мы получаем

$$f''(x) = 2c_2 + 2\cdot 3c_3x + \ldots + (n-1)nc_nx^{n-2} + \ldots; \qquad (13'')$$

подставляя затем в полученную формулу (13″) $x=0$, мы видим, что

$$2!\,c_2 = f''(0).$$

Аналогично, продифференцировав (13″) и затем подставив $x=0$, получаем

$$3!\,c_3 = f'''(0)$$

и, продолжая дальше таким же образом, мы найдем общую формулу для коэффициента c_n:

$$c_n = \frac{1}{n!}\,f^{(n)}(0),$$

где $f^{(n)}(0)$ представляет собой значение n-й производной от функции $f(x)$ при $x=0$. В результате получим ряд Тейлора

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2!}\,f''(0) + \frac{x^3}{3!}\,f'''(0) + \ldots. \qquad (14)$$

Пусть читатель в качестве упражнения проверит, что в примерах (4)–(11) коэффициенты степенных рядов составлены как раз по этому закону.

2. Формула Эйлера $\cos x + i \sin x = e^{ix}$. Одним из самых поразительных достижений Эйлера, полученных им на основе его формалистических манипуляций, является открытие тесной внутренней связи, существующей в области комплексного переменного между функциями синус и косинус, с одной стороны, и показательной функцией — с другой. Нужно заранее указать, что ни «доказательство» Эйлера, ни следующие далее доводы ни в какой мере не носят строгого характера; это — типичные для XVIII в. примеры формальных буквенных выкладок, пронизанных доверием к силе математического символизма.

Начнем с тождества Муавра, доказанного в главе II:

$$(\cos n\varphi + i \sin n\varphi) = (\cos \varphi + i \sin \varphi)^n.$$

Подстановка $\varphi = \frac{x}{n}$ приводит нас к соотношению

$$\cos x + i \sin x = \left(\cos \frac{x}{n} + i \sin \frac{x}{n}\right)^n.$$

Если x зафиксировано, то $\cos \frac{x}{n}$ будет мало отличаться от $\cos 0 = 1$ при неограниченном возрастании n; кроме того, так как

$$\frac{\sin \frac{x}{n}}{\frac{x}{n}} \to 1 \quad \text{при} \quad \frac{x}{n} \to 0$$

(см. стр. 329), то мы заключаем, что $\sin \frac{x}{n}$ асимптотически равен $\frac{x}{n}$. Поэтому можно считать более или менее естественным такой предельный переход:

$$\cos x + i \sin x = \lim \left(1 + \frac{ix}{n}\right)^n \quad \text{при} \quad n \to \infty.$$

Преобразуя правую часть этого равенства согласно формуле (стр. 473)

$$e^z = \lim \left(1 + \frac{z}{n}\right)^n \quad \text{при} \quad n \to \infty,$$

мы получим соотношение

$$\cos x + i \sin x = e^{ix}. \tag{15}$$

Это и есть результат, полученный Эйлером.

Мы можем вывести эту самую формулу и другим, тоже формалистическим путем — из разложения функции e^z:

$$e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots,$$

вместо z подставляя ix, где x — действительное число. Если мы вспомним, что последовательными степенями числа i являются числа i, -1, $-i$, $+1$ и т. д., периодически, то, собирая действительные и мнимые части, мы получим

$$e^{ix} = \left(1 - \frac{x^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \ldots\right) + i\left(x - \frac{x^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \ldots\right);$$

заменяя далее ряды в правой части их суммами $\cos x$ и $\sin x$, мы снова получаем формулу Эйлера.

Такое рассуждение отнюдь не является настоящим доказательством соотношения (15). Против нашего второго «вывода» можно возразить, что разложение в ряд для функции e^z было проведено в предположении, что z — действительное число; поэтому подстановка $z = ix$ должна быть оправдана дополнительными соображениями. Точно так же полноценность первого рассуждения уничтожается тем, что формула

$$e^z = \lim \left(1 + \frac{z}{n}\right)^n \quad \text{при} \quad n \to \infty$$

была раньше выведена только для действительных значений z.

Чтобы формула Эйлера из области чистого формализма перешла в область строгих математических истин, потребовалось развитие теории функций комплексного переменного — одного из величайших достижений XIX в. Многие другие глубокие проблемы стимулировали это далеко идущее развитие. Мы видели, например, что промежутки сходимости разложений различных функций в степенные ряды различны. Почему некоторые разложения сходятся всюду, т. е. для всех значений x, в то время как другие теряют смысл при $|x| > 1$?

Рассмотрим, например, геометрическую прогрессию (4), приведенную на стр. 498, которая сходится при $|x| < 1$. Левая часть этого равенства вполне осмысленна при $x = 1$, именно, равна $\frac{1}{1+1} = \frac{1}{2}$; в то же время ряд в правой части ведет себя очень странно: он принимает вид $1 - 1 + 1 - 1 + \ldots$

Этот последний ряд не является сходящимся, поскольку его частные суммы колеблются между 1 и 0. Это свидетельствует о том, что функция может порождать расходящийся ряд даже в том случае, если сама она не обнаруживает какой-либо иррегулярности. Правда, функция $\frac{1}{1+x}$ становится бесконечной при $x = -1$. И так как легко доказать, что сходимость степенного ряда в точке $x = a > 0$ влечет за собой сходимость в промежутке $-a < x < a$, то мы могли бы, пожалуй, усмотреть «объяснение» странного поведения нашего разложения в разрывности функции $\frac{1}{1+x}$ при $x = -1$. Однако рассмотрим теперь функцию $\frac{1}{1+x^2}$; она может быть разложена в ряд

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \ldots,$$

как мы убеждаемся, подставляя x^2 вместо x в формулу (4). Полученный ряд тоже сходится при $|x| < 1$; вместе с тем при $x = 1$ он снова приводит к ряду $1 - 1 + 1 - 1 + \ldots$, а при $|x| > 1$ он резко расходится, и однако же сама функция всюду ведет себя безупречно.

Оказывается, что полное объяснение этим явлениям возможно лишь

тогда, когда функции изучаются в области комплексных значений переменного x, охватывающей как действительные, так и мнимые его значения. Например, ряд для функции $\dfrac{1}{1+x^2}$ должен расходиться при $x = i$, так как знаменатель дроби при этом значении переменного равен нулю. Отсюда следует, что ряд должен расходиться при всех таких значениях x, что $|x| > |i| = 1$, поскольку можно доказать, что сходимость его для одного такого значения x повлекла бы за собой его сходимость при $x = i$. Таким образом, вопрос о сходимости рядов, которым полностью пренебрегали в период возникновения анализа, стал одним из главных факторов создания теории функций комплексного переменного.

3. Гармонический ряд и дзета-функция. Формула Эйлера, выражающая $\sin x$ в виде бесконечного произведения. Ряды, члены которых являются простыми комбинациями целых чисел, особенно интересны. В качестве примера рассмотрим «гармонический ряд»

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{n} + \ldots, \tag{16}$$

отличающийся от известного нам ряда, сумма которого равна $\ln 2$, только знаками членов, стоящих на четных местах.

Поставить вопрос, сходится ли этот ряд, все равно, что спросить себя, стремится ли к конечному пределу последовательность чисел

$$s_1, \quad s_2, \quad s_3, \quad \ldots,$$

где

$$s_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}. \tag{17}$$

Несмотря на то, что по мере продвижения по ряду (16) члены его приближаются к 0, легко увидеть, что ряд этот не сходится. Действительно, взяв достаточное количество членов, мы можем превысить любое положительное число; таким образом, s_n возрастает беспредельно, и, значит, ряд (16) «расходится к бесконечности». Чтобы в этом убедиться, заметим, что

$$s_2 = 1 + \frac{1}{2},$$

$$s_4 = s_2 + \left(\frac{1}{3} + \frac{1}{4}\right) > s_2 + \left(\frac{1}{4} + \frac{1}{4}\right) = 1 + \frac{2}{2},$$

$$s_8 = s_4 + \left(\frac{1}{5} + \ldots + \frac{1}{8}\right) > s_4 + \left(\frac{1}{8} + \ldots + \frac{1}{8}\right) = s_4 + \frac{1}{2} > 1 + \frac{3}{2},$$

и вообще,

$$s_{2^m} > 1 + \frac{m}{2}. \tag{18}$$

Таким образом, например, частные суммы s_{2^m} превышают 100, если только $m \geqslant 200$.

Если гармонический ряд расходится, то, с другой стороны, можно доказать, что ряд

$$1+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{4^s}+\ldots+\frac{1}{n^s}+\ldots \tag{19}$$

сходится при всяком значении s, большем чем 1, и сумма его, рассматриваемая как функция переменного s, есть так называемая дзета-функция:

$$\zeta(s)=\lim\left(1+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{4^s}+\ldots+\frac{1}{n^s}\right)\quad \text{при}\quad n\to\infty. \tag{20}$$

Эта функция, очевидно, определена только при $s>1$.

Существует важное соотношение между дзета-функцией и простыми числами, которое мы выведем, исходя из свойств геометрической прогрессии. Пусть p есть какое-нибудь простое число; тогда при $s\geqslant 1$

$$0<\frac{1}{p^s}<1,$$

так что

$$\frac{1}{1-\frac{1}{p^s}}=1+\frac{1}{p^s}+\frac{1}{p^{2s}}+\frac{1}{p^{3s}}+\ldots$$

Перемножим такого рода равенства, написанные для всех простых чисел $p=2,\ 3,\ 5,\ 7,\ \ldots$ (не задаваясь вопросом о законности такой операции). В левой части мы получим «бесконечное произведение»

$$\left(\frac{1}{1-\frac{1}{2^s}}\right)\cdot\left(\frac{1}{1-\frac{1}{3^s}}\right)\cdot\left(\frac{1}{1-\frac{1}{5^s}}\right)\ldots=\lim\left[\frac{1}{1-\frac{1}{p_1^s}}\ldots\frac{1}{1-\frac{1}{p_n^s}}\right]\ \text{при}\ \ n\to\infty;$$

в это же время в правой части мы получаем ряд

$$1+\frac{1}{2^s}+\frac{1}{3^s}+\ldots=\zeta(s),$$

в силу того обстоятельства, что каждое целое число, большее чем 1, может быть единственным образом представлено как произведение степеней различных простых чисел. Итак, нам удалось выразить дзета-функцию в виде произведения

$$\zeta(s)=\left(\frac{1}{1-\frac{1}{2^s}}\right)\cdot\left(\frac{1}{1-\frac{1}{3^s}}\right)\cdot\left(\frac{1}{1-\frac{1}{5^s}}\right)\ldots \tag{21}$$

Если бы существовало *только конечное* число простых чисел, скажем, $p_1,\ p_2,\ p_3,\ \ldots,\ p_r$, то произведение в правой части формулы (21) было бы обыкновенным конечным произведением и имело бы поэтому конечное значение даже при $s=1$. Однако мы видели, что дзета-ряд при $s=1$

$$\zeta(1)=1+\frac{1}{2}+\frac{1}{3}+\ldots$$

расходится, стремясь к бесконечности. Это рассуждение, которое легко превратить в строгое доказательство, показывает, что существует бесконечное множество простых чисел. Конечно, это доказательство гораздо

запутаннее и искусственнее, чем данное Евклидом (см. стр. 40). Но оно столь же привлекательно, как трудный подъем на вершину горы, которая могла бы быть достигнута с другой стороны по комфортабельной дороге.

С помощью бесконечных произведений, подобных тому, которое дается формулой (21), функции иногда выражаются так же удобно, как и с помощью бесконечных рядов.

Другое бесконечное произведение, открытие которого представляет собой еще одно из достижений Эйлера, относится к тригонометрической функции $\sin x$. Чтобы понять найденную Эйлером формулу, мы начнем со следующего замечания относительно многочленов. Если

$$f(x) = a_0 + a_1 x + \ldots + a_n x^n$$

есть многочлен степени n и имеет n различных нулей $x_1, \ldots, x_n$, то, как известно из алгебры, функция $f(x)$ может быть разложена на линейные множители

$$f(x) = a_n (x - x_1) \ldots (x - x_n)$$

(см. стр. 122). Вынося за скобку произведение $x_1 x_2 \ldots x_n$, мы можем написать

$$f(x) = C\left(1 - \frac{x}{x_1}\right)\left(1 - \frac{x}{x_2}\right) \ldots \left(1 - \frac{x}{x_n}\right),$$

где C — постоянная, равная a_0, что легко установить, положив $x = 0$. Далее возникает вопрос: возможно ли аналогичное разложение уже не для полиномов, а для более сложных функций $f(x)$? (В общем случае ответ не может быть утвердительным, в чем легко убедиться на примере показательной функции, которая вовсе не имеет нулей, поскольку $e^x \neq 0$ при любых значениях x.) Эйлер открыл, что для функции синус такое разложение возможно. Чтобы написать формулу в ее простейшем виде, мы рассмотрим не $\sin x$, а $\sin \pi x$. Последняя функция имеет нулями точки $x = 0, \pm 1, \pm 2, \pm 3, \ldots$, так как $\sin \pi n = 0$ при всех целых n; иных же нулей она не имеет никаких. Формула Эйлера устанавливает соотношение

$$\sin \pi x = \pi x \left(1 - \frac{x^2}{1^2}\right)\left(1 - \frac{x^2}{2^2}\right)\left(1 - \frac{x^2}{3^2}\right)\left(1 - \frac{x^2}{4^2}\right) \ldots \qquad (22)$$

Стоящее справа бесконечное произведение сходится при всех значениях x и является одной из красивейших формул математики. При $x = \frac{1}{2}$ формула дает

$$\sin \frac{\pi}{2} = 1 = \frac{\pi}{2}\left(1 - \frac{1}{2^2 \cdot 1^2}\right)\left(1 - \frac{1}{2^2 \cdot 2^2}\right)\left(1 - \frac{1}{2^2 \cdot 3^2}\right) \ldots$$

Если мы напишем

$$1 - \frac{1}{2^2 n^2} = \frac{(2n-1)(2n+1)}{2n \cdot 2n},$$

то после небольших преобразований получим произведение Уоллиса

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \frac{8}{7} \cdot \frac{8}{9} \ldots,$$

упомянутое на стр. 322.

За доказательствами всех этих соотношений мы вынуждены направить читателя к руководствам по анализу (см. также стр. 532–533).

*§4. Доказательство теоремы о простых числах на основе статистического метода

Применяя математические методы к изучению явлений природы, обычно удовлетворяются такими рассуждениями, в ходе которых цепь строгих логических доводов прерывается более или менее правдоподобными допущениями. И даже в чистой математике можно подчас встретить рассуждение, которое хотя и не обеспечивает строгого доказательства, но все же, несмотря на это, подсказывает правильное решение и дает направление, в котором можно искать это строгое доказательство.

Именно таков характер решения задачи о брахистохроне, данного Якобом Бернулли (см. стр. 404), а также очень многих других проблем раннего периода развития анализа.

Пользуясь процедурой, типичной для прикладной математики и в особенности для статистической механики, мы приведем сейчас одно рассуждение, которое сделает по меньшей мере правдоподобной справедливость знаменитого гауссова закона о распределении простых чисел. (Упомянутая процедура была подсказана одному из авторов специалистом по экспериментальной физике Густавом Герцем.) Эта теорема, рассмотренная с эмпирической точки зрения в дополнениях к главе I, утверждает, что число $A(n)$ простых чисел, не превышающих n, асимптотически равно $\frac{n}{\ln n}$:

$$A(n) \sim \frac{n}{\ln n}.$$

Под этим подразумевается то, что отношение $A(n) : \frac{n}{\ln n}$ стремится к пределу при стремлении n к бесконечности.

Допустим прежде всего, что *существует* математический закон распределения простых чисел, обладающий следующим свойством: при больших значениях n определенная выше функция $A(n)$ приблизительно равна интегралу $\int\limits_2^n W(x)\,dx$, где $W(x)$ можно назвать функцией, измеряющей «плотность» простых чисел. (В качестве нижнего предела интеграла мы выбрали число 2, так как при $x < 2$ имеем, очевидно, $A(x) = 0$.) Более точно, пусть x — растущая величина и Δx — другая растущая величина, но такая, что порядок возрастания x больше порядка возрастания Δx. (Например, можно принять, что $\Delta x = \sqrt{x}$.) Предположим далее, что распределение простых чисел настолько равномерно, что число простых чисел в промежутке от x до $x + \Delta x$ приближенно равно выражению вида $W(x)\Delta x$, и даже более того,

что функция $W(x)$ изменяется так плавно, что интеграл $\int\limits_2^n W(x)\,dx$ может быть заменен соответствующей интегральной «ступенчатой» суммой, не изменяя своего асимптотического значения. После этих предварительных замечаний мы подготовлены для того, чтобы начать наше рассуждение.

Мы уже доказали ранее (стр. 496), что при больших целых числах выражение $\ln(n!)$ асимптотически равно произведению $n \ln n$:

$$\ln(n!) \sim n \ln n.$$

Мы дадим сейчас другую асимптотическую формулу для $\ln(n!)$, выражающуюся через простые числа, а затем сравним оба выражения. Сосчитаем, сколько раз произвольное простое число p, меньшее, чем n, входит множителем в целое число $p! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$. Обозначим символом $[a]_p$ наибольшее целое число k такое, что a делится на p^k. Из того, что разложение любого целого числа на простые числа единственно, вытекает зависимость $[ab]_p = [a]_p + [b]_p$ при любых целых a и b. Отсюда следует

$$[n!]_p = [1]_p + [2]_p + [3]_p + \ldots + [n]_p.$$

В последовательности 1, 2, 3, ..., n члены, делящиеся на p^k, имеют вид p^k, $2p^k$, $3p^k$, ...; их число N_k при больших n приближенно равно $\frac{n}{p^k}$.

Из этих членов число M_k таких, которые делятся на p^k, но не делятся на более высокие степени p, равно разности $N_k - N_{k+1}$. Следовательно, имеем

$$\begin{aligned}[n!]_p &= M_1 + 2M_2 + 3M_3 + \ldots = \\ &= (N_1 - N_2) + 2(N_2 - N_3) + 3(N_3 - N_4) + \ldots = N_1 + N_2 + N_3 + \ldots = \\ &= \frac{n}{p} + \frac{n}{p^2} + \frac{n}{p^3} + \ldots = \frac{n}{p-1}.\end{aligned}$$

(Само собой разумеется, что эти равенства только приближенные.)

Отсюда следует, что при больших n число $n!$ приближенно выражается произведением всех выражений вида $p^{\frac{n}{p-1}}$ с ограничением $p < n$. Таким образом, мы получили формулу

$$\ln(n!) \sim \sum_{p<n} \frac{n}{p-1} \ln p.$$

Сравнивая полученное выражение с нашей прежней асимптотической формулой для $\ln(n!)$, мы находим:

$$\ln x \sim \sum_{p<x} \frac{\ln p}{p-1} \qquad (1)$$

(вместо n подставлено x).

Следующим — и решающим — шагом является нахождение асимптотического выражения для правой части соотношения (1). Если x очень

велико, можно промежуток от 2 до $x = n$ разделить на большое число r достаточно больших частных промежутков точками $2 = \xi_1, \xi_2, \ldots, \xi_r, \xi_{r+1} = x$ с соответственными приращениями $\Delta\xi_j = \xi_{j+1} - \xi_j$. В каждом частном промежутке могут находиться простые числа, и каждое простое число j-го частного промежутка приближенно равно числу ξ_j. В силу нашего предположения о функции $W(x)$, в j-м частном промежутке содержится приблизительно $W(\xi_j)\Delta\xi_j$ простых чисел; следовательно, сумма в правой части соотношения (1) приближенно равна выражению

$$\sum_{j=1}^{r+1} W(\xi_j)\,\frac{\ln \xi_j}{\xi_j - 1}\,\Delta\xi_j.$$

Заменив эту конечную сумму интегралом, к которому она приближается, мы получим формулу

$$\ln x \sim \int_2^x W(\xi)\frac{\ln \xi}{\xi - 1}\,d\xi. \tag{2}$$

Отсюда мы теперь определим неизвестную функцию $W(x)$. Если мы заменим значок $\sim$ знаком обыкновенного равенства и продифференцируем обе части по x, то в силу основной теоремы анализа можно написать

$$\frac{1}{x} = W(x)\frac{\ln x}{x-1}, \qquad W(x) = \frac{x-1}{x\ln x}. \tag{3}$$

В самом начале нашего рассуждения мы предположили, что $A(x)$ асимптотически равно интегралу

$$\int_2^x W(x)\,dx.$$

Итак, $A(x)$ приближенно равно интегралу

$$\int_2^x \frac{x-1}{x\ln x}\,dx. \tag{4}$$

Для того чтобы вычислить этот интеграл, заметим, что функция $f(x) = \dfrac{x}{\ln x}$ имеет следующую производную:

$$f'(x) = \frac{1}{\ln x} - \frac{1}{(\ln x)^2}.$$

При больших значениях x два выражения

$$\frac{1}{\ln x} - \frac{1}{(\ln x)^2} \quad \text{и} \quad \frac{1}{\ln x} - \frac{1}{x\ln x}$$

асимптотически равны, так как вторые члены в обоих выражениях много меньше первых. Следовательно, интеграл (4) будет асимптотически

равен интегралу

$$\int_2^x f'(x)\,dx = f(x) - f(2) = \frac{x}{\ln x} - \frac{2}{\ln 2},$$

так как две сравниваемые нами функции мало отличаются на всем промежутке интегрирования. При больших значениях x членом $\frac{2}{\ln 2}$ как постоянным можно пренебречь, и тогда мы приходим к окончательному результату

$$A(x) \sim \frac{x}{\ln x}.$$

Это и есть теорема о простых числах.

Мы не можем претендовать на то, чтобы предыдущее рассуждение рассматривалось как математическое доказательство, а не как индукция. Однако более глубокий анализ приводит к следующему заключению. Нетрудно оправдать каждый, с такою смелостью сделанный нами шаг, в частности, доказать справедливость асимптотической формулы суммы и интеграла, стоящих соответственно в правых частях соотношений (1) и (2), и, наконец, обосновать шаг, ведущий от соотношения (2) к соотношению (3). Гораздо труднее *доказать существование* гладкой функции «плотности» $W(x)$, которое мы постулировали в самом начале. Но раз это принято, то *оценка* самой функции $W(x)$ является делом сравнительно простым; таким образом, в задаче о распределении простых чисел наиболее трудным представляется доказательство существования «плотности» $W(x)$.

ПРИЛОЖЕНИЕ

Дополнительные замечания. Задачи и упражнения

Многие из следующих задач предназначены для более или менее подготовленного читателя. Они имеют в виду не столько развитие рутинной техники операций, сколько находчивости и инициативы.

Арифметика и алгебра

1. Откуда мы знаем, что, как говорится на стр. 81, никакая степень 10 не делится на 3? (См. стр. 53.)

2. Докажите, что принцип наименьшего целого числа вытекает как следствие из принципа математической индукции. (См. стр. 28.)

3. Применяя биномиальную теорему к разложению $(1+1)^n$, докажите, что $C_0^n + C_1^n + C_2^n + \ldots + C_n^n = 2^n$.

*4. Задумайте какое-нибудь число $z = abc\ldots$, составьте сумму его цифр $a+b+c+\ldots$, вычтите ее из z, вычеркните одну цифру из остатка и обозначьте через w сумму оставшихся цифр. Нельзя ли найти правило для того, чтобы определить вычеркнутую цифру, зная только значение w? (Будет не вполне определенный случай, если $w = 0$.) Как и многое другое из того, что связано со сравнениями, этот пример пригоден в качестве «развлекательной задачи с угадыванием».

5. Арифметической прогрессией первого порядка называется такая последовательность чисел a, $a+d$, $a+2d$, $a+3d$, …, что разность между следующим членом и предыдущим постоянна. Арифметическая прогрессия второго порядка есть такая числовая последовательность a_1, a_2, a_3, …, что разности $a_{i+1} - a_i$ образуют арифметическую прогрессию первого порядка. Вообще арифметической прогрессией порядка k называют последовательность, обладающую тем свойством, что разности рядом стоящих членов образуют арифметическую прогрессию $(k-1)$-го порядка. Проверьте, что квадраты натуральных чисел образуют арифметическую прогрессию 2-го порядка, и затем установите по математической индукции что k-е степени натуральных чисел образуют арифметическую прогрессию порядка k. Докажите, что всякая последовательность чисел a_n, где $a_n = c_0 + c_1 n + c_2 n^2 + \ldots + c_k n^k$ и все c — постоянные, есть арифметическая прогрессия порядка k. *Докажите

обратное утверждение для случая $k=2$, $k=3$, для любого k.

6. Докажите, что суммы n первых членов арифметической прогрессии порядка k образуют арифметическую прогрессию $(k+1)$-го порядка.

7. Сколько делителей у числа 10 296? (См. стр. 43.)

8. Пользуясь алгебраической формулой

$$(a^2+b^2)(c^2+d^2)=(ac-bd)^2+(ad+bc)^2,$$

докажите с помощью индукции, что любое целое число $r=a_1a_2\ldots a_n$ может быть представлено как сумма двух квадратов, если каждый из множителей a обладает тем же свойством. Проверьте это на примерах

$$r=160,\qquad r=1600,\qquad r=1300,\qquad r=625,$$

принимая во внимание, что

$$2=1^2+1^2,\qquad 5=1^2+2^2,\qquad 8=2^2+2^2\quad \text{и т. д.}$$

Если возможно, найдите два различных представления этих чисел как суммы двух квадратов.

9. Воспользуйтесь результатом предыдущего упражнения для того, чтобы по заданным пифагоровым тройкам чисел строить новые.

10. Составьте правила делимости (подобные приведенным на стр. 53) для систем счисления с основаниями 7, 11, 12.

11. Докажите: неравенство $r>s$ между двумя положительными рациональными числами $r=\frac{a}{b}$ и $s=\frac{c}{d}$ равносильно неравенству $ac-bd>0$.

12. Докажите: если r и s — положительные числа, причем $r<s$, то

$$r<\frac{r+s}{2}<s\quad \text{и}\quad \frac{2}{\left(\frac{1}{r}+\frac{1}{s}\right)^2}<2rs<(r+s)^2.$$

13. Предполагая, что z — какое угодно комплексное число, докажите с помощью индукции, что $z^n+\frac{1}{z^n}$ может быть представлено как полином степени n относительно $w=z+\frac{1}{z}$. (См. стр. 121.)

*14. Если положим ради краткости $E(\varphi)=\cos\varphi+i\sin\varphi$, то получим $[E(\varphi)]^m=E(m\varphi)$. Воспользовавшись этой формулой, а также формулой для суммы геометрической прогрессии (стр. 33; эта формула сохраняется и для комплексных величин), докажите, что

$$\sin\varphi+\sin 2\varphi+\sin 3\varphi+\ldots+\sin n\varphi=\frac{\cos\frac{\varphi}{2}-\cos\left(n+\frac{1}{2}\right)\varphi}{2\sin\frac{\varphi}{2}},$$

$$\frac{1}{2}+\cos\varphi+\cos 2\varphi+\cos 3\varphi+\ldots+\cos n\varphi=\frac{\sin\left(n+\frac{1}{2}\right)\varphi}{2\sin\frac{\varphi}{2}}.$$

15. Что получится из формулы упражнения 3 на стр. 36, если подставить $q = E(\varphi)$?

Аналитическая геометрия

Внимательное выполнение следующих упражнений, сопровождаемое чертежами и числовыми примерами, даст возможность овладеть элементами аналитической геометрии. Определения и простейшие факты из тригонометрии предполагаются известными.

Во многих случаях целесообразно представлять себе прямые или отрезки направленными от одной точки к другой. Под *направленной* прямой PQ (или *направленным* отрезком PQ) мы будем понимать прямую (или отрезок), направленную от P к Q. Если разъяснения отсутствуют, то предполагается, что направленная прямая l имеет безразлично какое, но вполне определенное направление; только в случае направленной оси x неизменно принимается, что она направлена от начала O к любой ее точке с положительной координатой x, и аналогично для оси y. О направленных прямых (или отрезках) мы будем говорить, что они параллельны, в том и только том случае, если они направлены одинаково. Направление направленного отрезка на направленной прямой может быть фиксировано знаком + или −, присоединяемым к расстоянию между конечными точками отрезка, смотря по тому, совпадает или не совпадает направление отрезка с направлением прямой. Целесообразно понятие «отрезок PQ» обобщить и на тот случай, когда точки P и Q совпадают; таким «отрезкам» приписывается длина нуль и не приписывается никакого направления.

16. Докажите: если $P_1(x_1,\, y_1)$ и $P_2(x_2,\, y_2)$ — какие-нибудь две точки, то координаты точки $P_0(x_0,\, y_0)$ — середины отрезка P_1P_2 — определяются формулами $x_0 = \dfrac{x_1 + x_2}{2}$, $y_0 = \dfrac{y_1 + y_2}{2}$. Установите более общее положение: если точки P_1 и P_2 различны, то та точка P_0 на направленной прямой P_1P_2, для которой отношение направленных отрезков $\dfrac{P_1P_0}{P_1P_2}$ равно k, имеет координаты

$$x_0 = (1-k)x_1 + kx_2, \qquad y_0 = (1-k)y_1 + ky_2.$$

(*Указание*: примите во внимание свойство пропорциональности отрезков при пересечении сторон угла параллельными прямыми.)

Таким образом, все точки на прямой P_1P_2 имеют координаты вида $x = \lambda_1 x_1 + \lambda_2 x_2$, $y = \lambda_1 y_1 + \lambda_2 y_2$, причем $\lambda_1 + \lambda_2 = 1$. При $\lambda_1 = 1$ и $\lambda_2 = 0$ получаются соответственно точки P_1 и P_2. Отрицательным значениям λ_1 соответствуют точки по ту сторону P_2, отрицательным значениям λ_2 — точки по ту сторону P_1, т. е. точки, лежащие на прямой P_1P_2 вне отрезка P_1P_2.

17. Охарактеризуйте подобным же образом положение точки на пря-

мой в зависимости от числовых значений k.

В такой же мере существенно использовать положительные и отрицательные числа для обозначения направленных вращений. По определению, в качестве положительного направления вращения избирается то, которое направленную ось x переводит в ось y после поворота на 90°. При обыкновенном расположении осей (когда ось x направлена вправо, а ось y — вверх) положительное вращение направлено против часовой стрелки. Мы определим теперь угол от направленной прямой l_1 к направленной прямой l_2 как угол, на который нужно повернуть прямую l_1, чтобы она стала параллельной прямой l_2. Разумеется, этот угол определен лишь с точностью до величин, кратных 360°. Так, угол между осью x и осью y равен 90° или -270° и т. д. Вместо «угол от l_1 до l_2» мы будем говорить короче: «угол между l_1 и l_2», учитывая, конечно, что «угол между l_1 и l_2» и «угол между l_2 и l_1» различаются знаком.

18. Пусть α обозначает угол между направленной осью x и направленной прямой l. Пусть P_1, P_2 — две точки на l и d — направленное расстояние от P_1 до P_2. Покажите, что

$$\cos\alpha = \frac{x_2 - x_1}{d}, \qquad \sin\alpha = \frac{y_2 - y_1}{d},$$
$$(x_2 - x_1)\sin\alpha = (y_2 - y_1)\cos\alpha.$$

Если прямая l не перпендикулярна к оси x, то *наклон* l определяется формулой

$$m = \operatorname{tg}\alpha = \frac{y_2 - y_1}{x_2 - x_1}.$$

Значение m не зависит от того, как направлена прямая l, так как $\operatorname{tg}\alpha = \operatorname{tg}(\alpha + 180^\circ)$, или, что равносильно,

$$\frac{y_1 - y_2}{x_1 - x_2} = \frac{y_2 - y_1}{x_2 - x_1}.$$

19. Докажите: наклон прямой равен нулю, положителен или отрицателен, смотря по тому, пойдет ли прямая, параллельная данной и проходящая через начало, по оси x или через первый и третий квадранты, или через второй и четвертый квадранты.

Мы условимся различать две «стороны» направленной прямой — положительную и отрицательную (менее удобно было бы говорить о положительной и отрицательной полуплоскостях). Пусть P — какая-нибудь точка, не лежащая на прямой l, и Q — основание перпендикуляра к l, проведенного через P. Тогда P лежит с положительной или с отрицательной стороны l, смотря по тому, будет ли угол между l и направленной прямой QP равен 90° или -90°.

Затем мы определим уравнение направленной прямой l. Через начало O проведем прямую m, перпендикулярную к l, и направим ее так, чтобы угол между m и l был равен 90°. Обозначим через β угол между направленной осью x и прямой m. Тогда $\alpha = 90^\circ + \beta$, $\sin\alpha = \cos\beta$, $\cos\alpha =$

$-\sin\beta$. Пусть $R(x_1, y_1)$ есть точка пересечения l и m. Обозначим через d направленное расстояние OR на направленной прямой m.

20. Покажите, что d положительно в том и только том случае, если начало O находится с отрицательной стороны l.

Мы имеем $x_1 = d\cos\beta$, $y_1 = d\sin\beta$ (см. упражнение 18). Отсюда следует $(x - x_1)\sin\alpha = (y - y_1)\cos\alpha$ или

$$(x - d\cos\beta)\cos\beta = -(y - d\sin\beta)\sin\beta,$$

что приводит окончательно к уравнению

$$x\cos\beta + y\sin\beta - d = 0.$$

Это — *нормальная форма* уравнения прямой l. Следует заметить, что это уравнение не зависит от направления прямой l, так как изменение ее направления повлекло бы за собой изменение знаков у всех членов в левой части, причем уравнение осталось бы тем же самым.

Умножая нормальное уравнение на произвольный множитель, мы получаем общую форму уравнения прямой линии

$$ax + by + c = 0.$$

Чтобы получить, обратно, из этой общей формы геометрически содержательную нормальную форму, придется умножить обе части уравнения на такой множитель, чтобы коэффициенты при x и y свелись к величинам вида $\cos\beta$ и $\sin\beta$, квадраты которых в сумме составляют 1. Таким множителем является $\dfrac{1}{\sqrt{a^2+b^2}}$; он придает уравнению нормальную форму

$$\frac{a}{\sqrt{a^2+b^2}}\,x + \frac{b}{\sqrt{a^2+b^2}}\,y + \frac{c}{\sqrt{a^2+b^2}} = 0;$$

здесь мы имеем

$$\frac{a}{\sqrt{a^2+b^2}} = \cos\beta, \qquad \frac{b}{\sqrt{a^2+b^2}} = \sin\beta, \qquad -\frac{c}{\sqrt{a^2+b^2}} = d.$$

21. Докажите, что: а) кроме $\dfrac{1}{\sqrt{a^2+b^2}}$ и $-\dfrac{1}{\sqrt{a^2+b^2}}$, не существует иных множителей, приводящих общее уравнение к нормальной форме; б) выбор того или иного множителя фиксирует направление прямой; в) после того как выбор того или иного множителя сделан, можно сказать, что начало O находится с положительной или с отрицательной стороны прямой или находится на самой прямой, смотря по тому, будет ли d отрицательным, положительным или нулем.

22. Докажите непосредственно, что прямая с наклоном m, проходящая через данную точку $P_0(x_0, y_0)$, представляется уравнением

$$y - y_0 = m(x - x_0), \qquad \text{или} \qquad y = mx + (y_0 - mx_0).$$

Докажите, что прямая, проходящая через две данные точки $P_1(x_1, y_1)$ и $P_2(x_2, y_2)$, имеет уравнение

$$(y_2 - y_1)(x - x_1) = (x_2 - x_1)(y - y_1).$$

Условимся: координата x точки пересечения прямой с осью x называется «отрезком на оси x»; аналогично относительно «отрезка на оси y».

23. Докажите, что, деля общее уравнение, полученное в упражнении 20, на надлежащим образом подобранное число, получим уравнение прямой «в отрезках на осях»

$$\frac{x}{a} + \frac{y}{b} = 1,$$

причем a и b — отрезки, которые прямая образует соответственно на оси x и оси y. Могут ли быть исключительные случаи?

24. Покажите, что в результате подобной же процедуры уравнение всякой прямой, не параллельной оси y, может быть «решено относительно y»:

$$y = mx + b.$$

(Если же прямая параллельна оси y, то ее уравнению можно придать вид $x = a$.)

25. Предположим, что $ax + by + c = 0$ и $a'x + b'y + c' = 0$ — уравнения двух данных прямых l и l'; пусть m и m' — соответственно их наклоны. Докажите, что l и l' параллельны или перпендикулярны, смотря потому, будет ли: ф) $m = m'$ или $mm' = -1$, б) $ab' - a'b = 0$ или $aa' + bb' = 0$. (Обратите внимание, что формулировка б) пригодна и для того случая, когда у прямой «нет никакого наклона», т. е. она параллельна оси y.)

26. Установите, что уравнение прямой, проходящей через точку $P_0(x_0, y_0)$ и параллельной данной прямой l с уравнением $ax + by + c = 0$, имеет вид $ax + by = ax_0 + by_0$. Установите, что если условие параллельности заменить условием перпендикулярности, то соответствующее уравнение примет вид $bx - ay = bx_0 - ay_0$. (Интересно заметить, что если уравнение l написано в нормальной форме, то в такой же форме получается и новое уравнение.)

27. Пусть уравнения $x\cos\beta + y\sin\beta - d = 0$ и $ax + by + c = 0$ представляют в нормальной форме и в общей форме одну и ту же прямую l. Докажите, что направленное расстояние h от l до некоторой точки $Q(u, v)$ дается формулой

$$h = u\cos\beta + v\sin\beta - d,$$

или же

$$h = \frac{au + bv + c}{\pm\sqrt{a^2 + b^2}},$$

и что h — положительное или отрицательное число, смотря по тому, лежит ли Q с положительной или с отрицательной стороны направленной прямой l (причем направление l фиксируется или углом β, или выбором знака при $\sqrt{a^2 + b^2}$). (*Указание*: напишите в нормальной форме уравнение прямой m, проходящей через Q и параллельной m, и затем определите расстояние между l и m.)

28. Предположим, что $l(x,\, y) = 0$ есть сокращенная запись уравнения $ax + by + c = 0$ некоторой прямой l; аналогично для $l'(x,\, y) = 0$. Пусть λ и λ' — постоянные числа, причем $\lambda + \lambda' = 1$. Докажите, что если прямые l и l' пересекаются в точке $P_0(x_0,\, y_0)$, то всякая прямая, проходящая через P_0, имеет уравнение вида

$$\lambda l(x,\, y) + \lambda' l'(x,\, y) = 0,$$

и наоборот, что всякая такая прямая однозначно определяется выбором пары значений λ и λ'. (*Указание*: в том и только том случае P_0 лежит на l, если $l(x_0,\, y_0) = ax_0 + by_0 + c = 0$.) Что представляет уравнение в случае, если l и l' параллельны? (Заметьте, что в условии $\lambda + \lambda' = 1$ нет никакой необходимости: оно служит только для того, чтобы каждой прямой, проходящей через P_0, соответствовало одно определенное уравнение.)

29. Воспользуйтесь результатами предыдущего упражнения для того, чтобы написать уравнение прямой, проходящей через точку пересечения P_0 прямых l и l' и еще через другую точку $P_1(x_1,\, y_1)$, не находя при этом координат P_0. (*Указание*: определите λ и λ' из условий $\lambda l(x_1,\, y_1) + \lambda' l'(x_1,\, y_1) = 0$, $\lambda + \lambda' = 1$. Сделайте проверку, определяя координаты P_0 (см. стр. 97) и устанавливая затем, что P_0 лежит на прямой, уравнение которой вы нашли.)

30. Докажите, что уравнения биссектрис углов, образованных пересекающимися прямыми l и l', имеют вид

$$\sqrt{a'^2 + b'^2}\, l(x,\, y) = \pm\sqrt{a^2 + b^2}\, l'(x,\, y).$$

(*Указание*: см. упражнение 27.) Что представляют эти уравнения, если прямые l и l' параллельны?

31. Двумя способами выведите уравнение прямой, проходящей через середину отрезка P_1P_2 и к нему перпендикулярной: а) напишите уравнение P_1P_2; найдите координаты середины P_0 отрезка P_1P_2; напишите уравнение прямой, проходящей через P_0 и перпендикулярной к P_1P_2; б) выразите в виде уравнения то условие, что расстояние (стр. 93) между P_1 и какой-нибудь точкой $P(x,\, y)$ искомой прямой равно расстоянию между P_2 и P; возведите обе части равенства в квадрат и сделайте упрощения.

32. Двумя способами выведите уравнение окружности, проходящей через три точки P_1, P_2, P_3, не лежащие на одной прямой: а) напишите уравнение перпендикуляров, проведенных к отрезкам P_1P_2 и P_2P_3 через их середины; найдите координаты центра как точки пересечения этих перпендикуляров, определите радиус как расстояние между центром и P_1; б) искомое уравнение должно иметь вид $x^2 + y^2 - 2ax - 2by = k$ (см. стр. 94). Так как каждая из данных точек лежит на окружности,

то мы должны иметь

$$x_1^2 + y_1^2 - 2ax_1 - 2by_1 = k,$$
$$x_2^2 + y_2^2 - 2ax_2 - 2by_2 = k,$$
$$x_3^2 + y_3^2 - 2ax_3 - 2by_3 = k,$$

так как точки лежат на кривой в том и только том случае, если ее координаты удовлетворяют уравнению кривой. Затем решите систему относительно a, b, k.

33. Напишем уравнение эллипса с большей осью $2p$, малой осью $2q$ и с фокусами $F(-e, 0)$ и $F'(e, 0)$, причем $e^2 = p^2 - q^2$. Воспользуемся расстояниями r и r' произвольной точки $P(x, y)$ кривой от F и F'. По определению эллипса $r + r' = 2p$. С помощью формулы для расстояния между точками (стр. 93) установите, что

$$r'^2 - r^2 = (x+e)^2 - (x-e)^2 = 4ex.$$

Так как

$$r'^2 - r^2 = (r'+r)(r'-r) = 2p(r'-r),$$

то отсюда выведите соотношение $r' - r = \dfrac{2ex}{p}$. Решая последнее уравнение совместно с уравнением $r + r' = 2p$, вы получите важные формулы

$$r = -\frac{e}{p}x + p, \qquad r' = \frac{e}{p}x + p.$$

Так как (опять по формуле расстояния) $r^2 = (x-e)^2 + y^2$, то можно будет приравнять полученное выражение для r^2 выражению $\left(-\dfrac{e}{p}x + p\right)^2$, полученному раньше, и тогда будем иметь

$$(x-e)^2 + y^2 = \left(-\frac{e}{p}x + p\right)^2.$$

Раскройте скобки, соберите члены, подставьте $p^2 - q^2$ вместо e^2 и сделайте упрощения. Приведите окончательно к виду

$$\frac{x^2}{p^2} + \frac{y^2}{q^2} = 1.$$

Сделайте аналогичные вычисления для гиперболы, определяя ее как геометрическое место точек P, для которых абсолютная величина разности $r - r'$ равна данному числу $2p$. Но в этом случае $e^2 = p^2 + q^2$.

34. Парабола определяется как геометрическое место точек, расстояние которых от данной прямой (директрисы) равно расстоянию от данной точки (фокуса). Выбрав в качестве директрисы прямую $x = -a$, а в качестве фокуса точку $F(a, 0)$, покажите, что уравнение параболы может быть написано в виде

$$y^2 = 4ax.$$

Геометрические построения

35. Докажите невозможность построения с помощью только циркуля и линейки чисел $\sqrt[3]{3}$, $\sqrt[3]{4}$, $\sqrt[3]{5}$. Докажите, что построение числа $\sqrt[3]{a}$ возможно только в том случае, если a есть куб рационального числа (см. стр. 155 и далее).

36. Найдите стороны правильных $3 \cdot 2^n$-угольников и $5 \cdot 2^n$-угольников. Дайте характеристику последовательно вводимых полей расширения.

37. Докажите невозможность трисекции с помощью только циркуля и линейки углов в 120° или 30°. (*Указание* для случая угла в 30°: мы приходим к уравнению $4z^3 - 3z = \cos 30^\circ = \frac{1}{2}\sqrt{3}$. Введя новую переменную $u = z\sqrt{3}$, вы получите уравнение, с которым рассуждайте так же, как на стр. 158.)

38. Докажите невозможность построения правильного 9-угольника.

39. Установите, что инверсия точки $P(x, y)$ в точку $P'(x', y')$ относительно окружности с центром в начале координат и радиусом r дается формулами

$$x' = \frac{xr}{x^2 + y^2}, \qquad y' = \frac{yr}{x^2 + y^2}.$$

Решите эти уравнения относительно x и y.

*40. Основываясь на упражнении 39, установите аналитически, что при инверсии окружности и прямые переходят в окружности и прямые. Проверьте, в частности, свойства а)–г) со стр. 163, а также преобразования, указанные на рис. 61.

41. Что станет с двумя семействами прямых $x = \text{const}$ и $y = \text{const}$, параллельных координатным осям, при инверсии относительно единичной окружности с центром в начале? Дайте ответ без помощи аналитической геометрии и с помощью аналитической геометрии.

42. Выполните построение Аполлония в простых случаях по вашему собственному выбору. Попробуйте найти решение в аналитической форме, как было указано на стр. 146.

Проективная и неевклидова геометрия

43. Найдите все значения двойного отношения λ четырех гармонических точек, если эти точки подвергаются всевозможным перестановкам. $\left(\textit{Ответ}\colon \lambda = -1,\ 2,\ \frac{1}{2}.\right)$

44. При каких расположениях четырех точек какие-нибудь два из шести значений двойного отношения на стр. 195 совпадают между собой? (*Ответ*: только при $\lambda = -1$ или $\lambda = 1$; имеется только одно мнимое значение λ, при котором $\lambda = \frac{1}{1-\lambda}$: ему соответствует «эквигармоническое» двойное отношение.)

45. Удостоверьтесь, что равенство двойного отношения $(ABCD)$ единице означает совпадение точек C и D.

46. Докажите утверждения, касающиеся двойного отношения плоскостей, приведенные на стр. 197.

47. Докажите: если точки P и P' взаимно обратны относительно окружности и диаметр AB коллинеарен с точками P и P', то четверка точек A, B, P, P' гармоническая.

48. Найдите координату четвертой гармонической точки относительно данных точек P_1, P_2, P_3. Что случится, если P_3 станет приближаться к середине отрезка P_1P_2? (См. стр. 205.)

*49. Попробуйте развить теорию конических сечений, исходя из сфер Данделена. В частности, докажите, что все они, за исключением окружностей, являются геометрическим местом точек, для которых расстояния от данной точки и от данной прямой находятся в постоянном отношении k. При $k > 1$ получается гипербола, при $k = 1$ — парабола, при $k < 1$ — эллипс. Прямая l получается как пересечение плоскости конического сечения с плоскостью того круга, по которому сфера Данделена соприкасается с конусом. (Именно по той причине, что круг приходится особо оговаривать как предельный случай, не совсем удобно принимать указанное свойство в качестве определения конических сечений, что, впрочем, делается довольно часто.)

50. Обсудите следующее положение: «коническое сечение, рассматриваемое одновременно как множество точек и как множество касательных прямых, само себе двойственно» (см. стр. 228).

*51. Попробуйте доказать теорему Дезарга, выполняя предельный переход от пространственной конфигурации, изображенной на рис. 73 (см. стр. 192).

*52. Сколько можно провести в пространстве прямых, пересекающихся с данными четырьмя прямыми? Как их можно характеризовать? (*Указание*: через три данные прямые проведите гиперболоид; см. стр. 233.)

*53. Возьмем в качестве круга Пуанкаре единичный круг в комплексной плоскости. Пусть z_1 и z_2 — какие-то две точки внутри этого круга, а w_1 и w_2 — точки пересечения с окружностью «прямой линии», проходящей через z_1 и z_2. Тогда двойное отношение $\dfrac{z_1 - w_1}{z_1 - w_2} : \dfrac{z_2 - w_1}{z_2 - w_2}$, в соответствии с упражнением 8 на стр. 118, имеет действительное значение; докажите это. По определению, его логарифм есть гиперболическое расстояние между z_1 и z_2.

*54. С помощью инверсии преобразуйте круг Пуанкаре в верхнюю полуплоскость. Исследуйте эту полуплоскость как модель Пуанкаре и непосредственно, исходя из преобразования инверсии.

Топология

55. Проверьте формулу Эйлера на пяти правильных многогранниках и на других многогранниках. Сделайте схемы соответствующих разверток.

56. При доказательстве формулы Эйлера (стр. 256) нам приходится путем последовательного выполнения двух основных операций редуцировать произвольную сетку треугольников к сетке, состоящей только из одного треугольника, и тогда получаем $V - E + F = 3 - 3 + 1 = 1$. Почему мы можем быть уверены, что в конечном результате не окажется двух треугольников без общих вершин, и тогда было бы $V - E + F = 6 - 6 + 2 = 2$? (*Указание*: можно с самого начала исходить из предположения, что сетка треугольников *связная*, т. е. что по ребрам (сторонам) можно пройти от любой вершины к любой. Докажите, что это свойство не теряется при выполнении каждой из основных операций.)

57. Мы допустили при редукции сетки треугольников только две основные операции. Но не могло ли бы случиться на какой-то стадии редукции, что у нас окажется треугольник, имеющий только одну общую вершину с прочими треугольниками сетки? (Постройте пример.) В таком случае потребовалась бы еще третья операция: удаление двух вершин, трех ребер и одной грани. Как это отразилось бы на доказательстве?

58. Можно ли вокруг палки обернуть три раза широкую резиновую ленту так, чтобы она всюду лежала плотно, т. е. не делала бы складок? (Конечно, лента должна где-то сама себя пересекать.)

59. Установите, что после удаления центральной точки круговой диск допускает непрерывное преобразование в самого себя без неподвижных точек.

*60. Преобразование, переводящее каждую точку диска на единицу в определенном, одном и том же, направлении, очевидно, не обладает неподвижными точками. Конечно, это не есть преобразование в себя, так как некоторые точки диска после преобразования окажутся вне диска. Почему в этом случае рассуждение, приведенное на стр. 272 (основанное на преобразовании $P \to P^*$), уже не годится?

61. Допустим, что внутренняя сторона тора выкрашена в белую краску, а внешняя — в черную. Можно ли, сделав маленькую дырочку в поверхности, деформировав ее и затем запечатав дырочку опять, вывернуть тор «наизнанку» — так, чтобы внутренняя сторона была черная, а внешняя — белая?

*62. Установите, что в трехмерном пространстве не существует «проблемы четырех красок»: каково бы ни было число n, всегда можно n тел расположить так, чтобы каждое из них имело общую поверхность с каждым.

*63. На торе или на плоскости с граничной идентификацией (рис. 143)

сделайте карту из семи областей, из которых каждая имела бы общую границу с каждой (см. стр. 267).

64. Четырехмерный тетраэдр, изображенный на рис. 118, состоит из пяти точек a, b, c, d, e, причем каждая связана отрезком с каждой. Даже если бы было позволено искривлять эти отрезки, всю фигуру нельзя было бы уместить в плоскости таким образом, чтобы соединяющие линии не пересекались. Другая конфигурация, содержащая *девять* соединяющих линий, которые нельзя провести без пересечения, составляется из шести точек a, b, c, a', b', c', причем каждая из точек a, b, c должна быть соединена с каждой из точек a', b', c'. Проверьте эти утверждения экспериментально и затем попробуйте найти доказательство, основанное на теореме Жордана. (Доказано, что любая конфигурация точек и линий, которую нельзя уместить в плоскости без пересечений, непременно содержит как часть одну из двух указанных здесь конфигураций.)

65. Рассмотрите конфигурацию, составленную из 6 ребер трехмерного тетраэдра с добавлением отрезка, связывающего середины двух противоположных ребер. (Два ребра тетраэдра считаются противоположными, если у них нет общей вершины.) Установите, что эта конфигурация эквивалентна одной из описанных в предыдущем упражнении.

*66. Пусть p, q, r обозначают три горизонтальные черты в букве E. Эта буква после перемещения дает другое E с горизонтальными чертами p', q', r'. Можно ли связать p с p', q с q', r с r' линиями, которые не пересекались бы взаимно и не пересекали бы ни одного из двух E?

67. Если мы обходим вокруг квадрата, то меняем направление четыре раза, всякий раз на $90°$, а всего — на $\Delta = 360°$. Если обходим вокруг треугольника, то, как известно из элементарной геометрии, и в этом случае общее изменение направления составляет $\Delta = 360°$.

Докажите, что в случае любого простого замкнутого многоугольника C получается $\Delta = 360°$. (*Указание*: разбейте внутренность C на треугольники, затем удаляйте граничные отрезки, как на стр. 259. Обозначим последовательно образующиеся границы через B_1, B_2, B_3, $\ldots$, B_n. Тогда $B_1 = C$, а B_n есть треугольник. Предполагая, что изменение Δ_i соответствует границе B_i, докажите, что $\Delta_i = \Delta_{i-1}$.)

*68. Пусть C — простой замкнутый контур, обладающий во всех точках касательным вектором с непрерывно меняющимся направлением, и пусть Δ есть общее изменение направления при обходе контура. Докажите, что и в этом случае $\Delta = 360°$. (*Указание*: пусть p_0, p_1, p_2, $\ldots$, p_n, p_0 — точки на контуре C, разбивающие C на маленькие, почти прямолинейные отрезки. Пусть контур C составлен из прямолинейных отрезков p_0p_1, p_1p_2, $\ldots$, $p_{i-1}p_i$ и из первоначальных дуг p_ip_{i+1}, $\ldots$, p_np_0. Тогда $C_0 = C$, а C_n есть многоугольник. Докажите, что $\Delta_i = \Delta_{i+1}$, и воспользуйтесь результатом предыдущего упражнения.) Справедливо ли это утверждение для гипоциклоиды, изображенной на рис. 55?

69. Покажите, что если на диаграмме бутылки Клейна (см. стр. 282) все четыре стрелки направить одинаково (по часовой стрелке), то получится поверхность, эквивалентная сфере, у которой односвязный кусок поверхности заменен кросс-кэпом. (Эта поверхность топологически эквивалентна также расширенной проективной плоскости.)

70. Бутылка Клейна, изображенная на рис. 142, может быть разрезана плоскостью на две симметрически расположенные части. Покажите, что каждая из этих частей есть лента Мёбиуса.

*71. В ленте Мёбиуса (см. рис. 139) отождествляются два конца каждого поперечного отрезка. Убедитесь, что результат топологически эквивалентен бутылке Клейна.

Все возможные пары точек на прямолинейном отрезке, взятых в определенном порядке (две точки могут и совпадать), образуют квадрат в следующем смысле. Если точки фиксируются их расстояниями x, y от одного из концов A, то пара чисел (x, y) может быть рассматриваема как прямоугольные координаты некоторой точки квадрата.

Все возможные пары точек на прямолинейном отрезке, взятых независимо от порядка (т. е. пара (x, y) и пара (y, x) рассматриваются как одинаковые), образуют поверхность S, топологически эквивалентную квадрату. Чтобы убедиться в этом, будем считать первой ту точку, которая ближе к концу A, если $x \neq y$. Тогда S состоит из всех пар (x, y), где или $x < y$ или $x = y$. В плоскости прямоугольных координат получается треугольник с вершинами (0, 0), (0, 1), (1, 1).

*72. Какая поверхность получается из множества пар точек, взятых в определенном порядке: первая — на прямолинейном отрезке, вторая — на окружности? (*Ответ*: цилиндр).

73. Какая поверхность получается из множества пар точек, взятых в определенном порядке, причем обе точки берутся на окружности? (*Ответ*: тор).

*74. Какая поверхность получается из множества пар точек, взятых независимо от порядка на окружности? (*Ответ*: лента Мёбиуса).

75. Вот правила игры с монетами (одинаковых размеров) на большом круглом столе. A и B кладут монеты на стол по очереди. Монеты не должны касаться друг друга; их можно класть на столе как угодно, лишь бы они не перекрывались и не выступали за край стола. Раз монета положена, двигать ее уже нельзя. Рано или поздно стол покроется монетами таким образом, что для новой монеты места уже не найдется. Выигрывает тот, кто положит монету последним. Докажите, что, как бы ни играл B, если только A начинает игру, он может быть уверен в выигрыше — лишь бы играл правильно.

76. Убедитесь, что в случае, если стол в предыдущем упражнении имеет форму, показанную на рис. 125, *б*, то B всегда имеет возможность выиграть.

Функции, пределы, непрерывность

77. Разложите в непрерывную дробь отношение $OB : AB$ со стр. 143.

78. Докажите, что последовательность $a_0 = \sqrt{2}$, $a_{n+1} = \sqrt{2+\sqrt{a_n}}$, монотонно возрастает, ограничена числом $B = 2$ и, значит, имеет предел. Докажите, что этот предел не может быть отличен от 2 (см. стр. 316 и 347).

*79. Попробуйте доказать посредством рассуждений, подобных тем, какие были приведены на стр. 339 и далее, что, какова бы ни была гладкая замкнутая кривая, всегда можно начертить квадрат, стороны которого будут касаться кривой.

80. Функция $u = f(x)$ называется *вогнутой*, если середина отрезка, соединяющего две любые точки соответствующего графика, лежит выше самого графика. Например, функция $u = e^x$ вогнутая (рис. 278), тогда как функция $u = \log x$ (рис. 277) — не вогнутая.

Докажите, что функция $u = f(x)$ вогнута в том и только том случае, если

$$\frac{f(x_1)+f(x_2)}{2} \geqslant f\left(\frac{x_1+x_2}{2}\right),$$

причем равенство допускается только при $x_1 = x_2$.

*81. Докажите, что в случае вогнутой функции оправдывается и более общее неравенство

$$\lambda_1 f(x_1) + \lambda_2 f(x_2) \geqslant f(\lambda_1 x_1 + \lambda_2 x_2),$$

где λ_1, λ_2 — две постоянные, подчиненные ограничениям $\lambda_1 + \lambda_2 = 1$, $\lambda_1 \geqslant 0$, $\lambda_2 \geqslant 0$. Это равносильно утверждению, что ни одна из точек отрезка, соединяющего две произвольные точки графика, не лежит ниже кривой.

82. Пользуясь условием упражнения 80, докажите, что функции $u = \sqrt{1+x^2}$ и $u = \dfrac{1}{x}$ (при $x > 0$) вогнутые, т. е. что при положительных значениях x_1 и x_2

$$\frac{\sqrt{1+x_1^2}+\sqrt{1+x_2^2}}{2} \geqslant \sqrt{1+\left(\frac{x_1+x_2}{2}\right)^2},$$

$$\frac{1}{2}\left(\frac{1}{x_1}+\frac{1}{x_2}\right) \geqslant \frac{2}{x_1+x_2}.$$

83. Докажите то же для $u = x^2$, $u = x^n$ при $x > 0$; для $u = \sin x$ при $\pi \leqslant x \leqslant 2\pi$; для $u = \operatorname{tg} x$ при $0 \leqslant x < \dfrac{\pi}{2}$ и для $u = -\sqrt{1-x^2}$ при $|x| \leqslant 1$.

Максимумы и минимумы

84. Найдите кратчайший путь от точки P к точке Q на рис. 178, если требуется подойти по очереди n раз к каждой из двух данных прямых (см. стр. 353).

85. Найдите кратчайший путь от точки P к точке Q внутри остроугольного треугольника, если требуется подойти к каждой стороне в данном порядке (см. стр. 353).

86. Наметьте линии уровня и удостоверьтесь в существовании по меньшей мере двух седловых точек на поверхности, расположенной над трехсвязной областью, границы которой находятся на одном и том же уровне (см. стр. 367). И здесь нужно сделать исключение для случая, когда касательная плоскость к поверхности горизонтальна вдоль некоторой кривой.

87. Исходя из двух произвольных положительных рациональных чисел a_0, b_0, одну за другой постройте пары $a_{n+1} = \sqrt{a_n b_n}$, $b_{n+1} = \frac{1}{2}(a_n + b_n)$. Докажите, что они образуют последовательность вложенных интервалов. (Предел этой последовательности при $n \to \infty$ есть так называемое арифметико-геометрическое среднее чисел a_0, b_0, игравшее большую роль в ранних исследованиях Гаусса.)

88. Найдите длину всего графика на рис. 219 и сравните с суммой длин двух диагоналей квадрата.

*89. Исследуйте, при каких условиях, наложенных на точки A_1, A_2, A_3, A_4 получается схема рис. 216 и при каких — схема рис. 218.

*90. Найдите такие расположения пяти точек, для которых существовали бы различные минимальные системы путей, удовлетворяющие угловым условиям. Некоторые из этих систем будут соответствовать относительным минимумам (см. стр. 364).

91. Докажите неравенство Шварца

$$(a_1b_1 + \ldots + a_nb_n)^2 \leqslant (a_1^2 + \ldots + a_n^2)(b_1^2 + \ldots + b_n^2),$$

справедливое для каких угодно a_i и b_i; докажите, что знак равенства возможен только при условии пропорциональности между числами a_i и b_i. (*Указание*: обобщите алгебраическую формулу, приведенную в упражнении 8.)

*92. Исходя из n положительных чисел $x_1, \ldots, x_n$, построим выражения s_k, определяемые формулами

$$s_k = \frac{x_1x_2\ldots x_k + \ldots}{C_n^k},$$

причем в числителе стоит сумма всевозможных произведений, составленных из всех сочетаний n чисел по k. Докажите, что

$$\sqrt[k+1]{s_{k+1}} \leqslant \sqrt[k]{s_k}$$

и что знак равенства возможен только в случае равенства всех чисел x_i.

93. При $n = 3$ эти неравенства сводятся к следующим:

$$\sqrt[3]{abc} \leqslant \sqrt{\frac{ab + bc + ca}{3}} \leqslant \frac{a + b + c}{3}.$$

Какие отсюда вытекают экстремальные свойства куба?

*94. Найдите дугу кривой минимальной длины, соединяющую две точки A, B и вместе с прямолинейным отрезком AB ограничивающую наперед заданную площадь. (*Ответ*: дуга должна быть круговая.)

*95. Даны два отрезка AB и $A'B'$. Найдите дуги кривых, соединяющие A с B и A' с B', ограничивающие вместе с отрезками данную площадь и обладающие наименьшей суммой длин. (*Ответ*: дуги должны быть круговыми, с одинаковыми радиусами.)

*96. Тот же вопрос — при каком угодно числе отрезков AB, $A'B'$ и т. д.

*97. Даны две прямые, пересекающиеся в точке O. Найдите на каждой из них по точке A и B и затем соедините эти точки кривой линией таким образом, чтобы при заданной площади, ограниченной кривой и обеими прямыми, длина дуги была минимальной. (*Ответ*: дуга должна быть круговой и перпендикулярной к обеим прямым.)

*98. Тот же вопрос со следующим видоизменением: обратить в минимум требуется не длину кривой AB, а весь периметр фигуры, т. е. сумму дуги AB и отрезков OA и OB. (*Ответ*: дуга — по-прежнему круговая, но выпячивается наружу, касаясь отрезков в их концах.)

*99. Обобщите эту проблему на случай нескольких угловых секторов.

*100. Установите, что «почти плоские» поверхности на рис. 240 не являются в точности плоскими, кроме стабилизирующей поверхности в центре куба. (*Замечание*: определить эти поверхности аналитически представляет заманчивую, еще не решенную проблему. То же относится и к поверхностям на рис. 251. Что касается рис. 258, то здесь в самом деле имеется 12 симметрических плоскостей, образующих по диагоналям углы в $120°$.)

Некоторые дополнительные предложения по поводу опытов с мыльными пленками. Сделайте опыты, указанные на рис. 256 и 257, при числе стержней, большем трех. Изучите, что происходит в предельных случаях, когда объем воздуха становится все меньше. Экспериментируйте с непараллельными плоскостями и другими поверхностями. Раздувайте центральный кубик на рис. 258, пока он не наполнит весь большой куб и не выпятится за пределы граней; потом выдувайте из него воздух, пытаясь обратить процесс.

*101. Найдите два равносторонних треугольника с данной суммой периметров и с минимальной суммой площадей. (*Ответ*: треугольники должны быть конгруэнтны. Воспользуйтесь методами дифференциального исчисления.)

*102. Найдите два треугольника с данной суммой периметров и максимальной суммой площадей. (*Ответ*: один треугольник вырождается в точку, другой должен быть равносторонним.)

*103. Найдите два треугольника с данной суммой площадей и минимальной суммой периметров.

*104. Найдите два равносторонних треугольника с данной суммой площадей и максимальной суммой периметров.

Дифференциальное и интегральное исчисления

105. Составьте производные от функций $\sqrt{1+x}$, $\sqrt{1+x^2}$, $\sqrt{\frac{x+1}{x-1}}$, исходя непосредственно из определения, затем преобразовывая разностное отношение таким образом, чтобы не представило труда вычислить предел при $x \to x_1$ (см. стр. 442–444).

106. Докажите, что функция $y = e^{-\frac{1}{x^2}}$ с дополнительным условием $y = 0$ при $x = 0$ имеет производные всех порядков, равные нулю, в точке $x = 0$.

107. Установите, что функция упражнения 106 не разлагается в ряд Тейлора в точке $x = 0$ (см. стр. 500).

108. Найдите точки перегиба $(f''(x) = 0)$ кривых

$$y = e^{-x^2} \text{ и } y = xe^{-x^2}.$$

109. Покажите, что если $f''(x)$ — полином с n различными корнями $x_1, x_2, \ldots, x_n$, то

$$\frac{f'(x)}{f(x)} = \sum_{i=1}^{n} \frac{1}{x - x_i}.$$

*110. Исходя из определения интеграла как предела суммы, докажите, что при $n \to \infty$

$$n\left(\frac{1}{1^2+n^2} + \frac{1}{2^2+n^2} + \ldots + \frac{1}{n^2+n^2}\right) \to \frac{\pi}{4}.$$

*111. Таким же образом докажите, что

$$\frac{b}{n}\left(\sin\frac{b}{n} + \sin\frac{2b}{n} + \ldots + \sin\frac{nb}{n}\right) \to \cos b - 1.$$

112. Нарисуйте рис. 276 на клетчатой бумаге в крупном масштабе и затем, подсчитывая маленькие квадратики, попадающие в заштрихованную область, найдите приближенное значение π.

113. Воспользуйтесь формулой (7) на стр. 465, чтобы вычислить π с погрешностью, не превышающей 0,01.

114. Докажите, что $e^{\pi i} = -1$ (см. стр. 502).

115. Данная замкнутая кривая увеличивается, расширяясь в отношении $1 : x$. Пусть $L(x)$ и $A(x)$ обозначают длину расширенной кривой и ограниченную ею площадь. Покажите, что $\frac{L(x)}{A(x)} \to 0$ при $x \to \infty$ и что даже $\frac{L(x)}{A(x)^k} \to 0$ при $x \to \infty$, если $k > \frac{1}{2}$. Проверьте это для окружности,

квадрата и *эллипса. (Площадь — более высокого порядка возрастания, чем длина кривой. См. стр. 493 и дальше.)

116. Показательная функция часто встречается в следующих комбинациях:

$$u = \operatorname{sh} x = \frac{1}{2}(e^x - e^{-x}), \qquad v = \operatorname{ch} x = \frac{1}{2}(e^x + e^{-x}),$$

$$w = \operatorname{th} x = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

называемых соответственно *гиперболическим синусом, гиперболическим косинусом и гиперболическим тангенсом.* Эти функции обладают многими свойствами, напоминающими свойства тригонометрических функций. Они связаны с гиперболой $u^2 - v^2 = 1$ так же, как тригонометрические функции $u = \cos x$ и $v = \sin x$ связаны с окружностью $u^2 + v^2 = 1$. Читателю предлагается проверить следующие формулы и сопоставить их с тригонометрическими формулами:

$$\frac{d(\operatorname{sh} x)}{dx} = \operatorname{ch} x, \qquad \frac{d(\operatorname{ch} x)}{dx} = \operatorname{sh} x, \qquad \frac{d(\operatorname{th} x)}{dx} = \frac{1}{\operatorname{ch}^2 x},$$

$$\operatorname{sh}(x + x') = \operatorname{sh} x \cdot \operatorname{ch} x' + \operatorname{ch} x \cdot \operatorname{sh} x',$$

$$\operatorname{ch}(x + x') = \operatorname{ch} x \cdot \operatorname{ch} x' + \operatorname{sh} x \cdot \operatorname{sh} x'.$$

Обратные функции таковы:

$$x = \operatorname{Arsh} u = \ln(u + \sqrt{u^2 + 1}),$$

$$x = \operatorname{Arch} v = \ln(v + \sqrt{v^2 - 1}) \quad (v \geqslant 1),$$

$$x = \operatorname{Arth} w = \frac{1}{2} \ln \frac{1 + w}{1 - w} \quad (|w| < 1).$$

Их производные имеют вид

$$\frac{d(\operatorname{Arsh}\ u)}{dx} = \frac{1}{\sqrt{1 + u^2}}, \qquad \frac{d(\operatorname{Arch}\ v)}{dx} = \frac{1}{\sqrt{v^2 - 1}},$$

$$\frac{d(\operatorname{Arth}\ w)}{dx} = \frac{1}{1 - w^2} \quad (|w| < 1).$$

117. Уясните себе аналогию между гиперболическими и тригонометрическими функциями на основе формулы Эйлера.

*118. Выведите простые формулы для сумм

$$\operatorname{sh} x + \operatorname{sh} 2x + \ldots + \operatorname{sh} nx$$

и

$$\frac{1}{2} + \operatorname{ch} x + \operatorname{ch} 2x + \ldots + \operatorname{ch} nx$$

аналогично формулам, выведенным в упражнении 14 в случае тригонометрических функций.

Техника интегрирования

Теорема, доказанная на стр. 462, сводит проблему интегрирования функции $f(x)$ в пределах от a до b к нахождению функции $G(x)$, первообразной по отношению к функции $f(x)$. Интеграл тогда просто равен разности $G(b) - G(a)$.

Для таких первообразных функций (определяемых с точностью до постоянного слагаемого) употребительно наименование «неопределенный интеграл» и чрезвычайно удобное обозначение

$$G(x) = \int f(x)\,dx,$$

без обозначения пределов интегрирования. (Это обозначение может несколько дезориентировать начинающего: см. замечания на стр. 462.)

Из каждой формулы дифференцирования легко получить, путем ее обращения, некоторую формулу неопределенного интегрирования. К этой, несколько эмпирической, процедуре мы здесь добавим два важных правила, которые по существу представляют собой не что иное, как обращение правил дифференцирования сложной функции и произведения двух функций. В их интегральной форме их называют правилами *интегрирования посредством подстановки* и *интегрирования «по частям»*.

A) Первое правило вытекает из формулы дифференцирования сложной функции

$$H(u) = G(x),$$

где функции

$$x = \psi(u) \quad \text{и} \quad u = \varphi(x)$$

предполагаются взаимно обратными в рассматриваемой области.

В таком случае мы имеем

$$H'(u) = G'(x)\,\psi'(u).$$

Полагая

$$G'(x) = f(x),$$

мы можем написать

$$G(x) = \int f(x)\,dx$$

и также

$$G'(x)\,\psi'(u) = f(x)\,\psi'(u),$$

а это вследствие предыдущей формулы для $H'(u)$ равносильно

$$H(u) = \int f[\psi(u)]\,\psi'(u)\,du.$$

Итак, принимая во внимание, что $H(u) = G(x)$, мы получаем

$$\int f(x)\,dx = \int f[\psi(u)]\,\psi'(u)\,du. \qquad \text{(I)}$$

Будучи записано в обозначениях Лейбница (см. стр. 457), это правило принимает практически очень удобный вид

$$\int f(x)\,dx = \int f(x)\,\frac{dx}{du}\,du;$$

оказывается, что мы не сделаем ошибки, если символ dx заменим символом $\frac{dx}{du}\,du$ — так, как будто бы dx и du были числами, а $\frac{dx}{du}$ — их отношением.

Проиллюстрируем полезность формулы (I) несколькими примерами.

а) $J = \int \frac{1}{u \ln u}\,du$. Станем читать формулу (I) справа налево, полагая в ней $x = \ln u = \psi(u)$. Тогда получим $\psi'(u) = \frac{1}{u}$, $f(x) = \frac{1}{x}$, так что

$$J = \int \frac{dx}{x} = \ln x,$$

или

$$\int \frac{du}{u \ln u} = \ln \ln u.$$

Результат можно проверить посредством дифференцирования; мы получаем

$$\frac{1}{u \ln u} = \frac{d}{du}(\ln \ln u).$$

б) $J = \int \operatorname{ctg} u\,du = \int \frac{\cos u}{\sin u}\,du$. Полагая $x = \sin u = \psi(u)$, мы имеем

$$\psi'(u) = \cos u, \qquad f(x) = x,$$

откуда следует

$$J = \int \frac{dx}{x} \ln x,$$

или

$$\int \operatorname{ctg} u\,du = \ln \sin u.$$

И этот результат проверяется дифференцированием.

в) Допустим, что задан интеграл более общего вида

$$J = \int \frac{\psi'(u)}{\psi(u)}\,du;$$

положив $x = \psi(u)$, $f(x) = x$, мы найдем:

$$J = \int \frac{dx}{x} = \ln x = \ln \psi(u).$$

г) $J = \int \sin x \cos x\,dx$. Полагаем $\sin x = u$, $\cos x = \frac{du}{dx}$. Тогда

$$J = \int u\,\frac{du}{dx}\,dx = \int u\,du = \frac{u^2}{2} = \frac{1}{2}\sin^2 x.$$

д) $J = \int \frac{\ln u}{u}\, du$. Полагаем $\ln u = x$, $\frac{1}{u} = \frac{dx}{du}$. Тогда

$$J = \int x \frac{dx}{du}\, du = \int x\, dx = \frac{x^2}{2} = \frac{1}{2}(\ln u)^2.$$

В следующих примерах мы используем формулу (I), считая ее слева направо.

е) $J = \int \frac{dx}{\sqrt{x}}$. Полагаем $\sqrt{x} = u$. Тогда $x = u^2$ и $\frac{dx}{du} = 2u$. Поэтому

$$J = \int \frac{1}{u} \cdot 2u\, du = 2u = 2\sqrt{x}.$$

ж) С помощью подстановки $x = au$, где a — постоянная, получаем

$$\int \frac{dx}{a^2 + x^2} = \int \frac{dx}{du} \cdot \frac{1}{a^2} \cdot \frac{1}{1+u^2}\, du = \int \frac{1}{a} \frac{du}{1+u^2} = \frac{1}{a} \cdot \operatorname{arctg} \frac{x}{a}.$$

з) $J = \int \sqrt{1 - x^2}\, dx$. Полагаем $x = \cos u$, $\frac{dx}{du} = -\sin u$. В таком случае

$$J = -\int \sin^2 u\, du = -\int \frac{1 - \cos 2u}{2}\, du = -\frac{u}{2} + \frac{\sin 2u}{4}.$$

Принимая во внимание, что

$$\sin 2u = 2 \sin u \cos u = 2 \cos u \sqrt{1 - \cos^2 u},$$

приходим к формуле

$$J = -\frac{1}{2} \arccos x + \frac{1}{2} x\sqrt{1 - x^2}.$$

Вычислите следующие интегралы и проверьте результаты посредством дифференцирования:

119. $\int \frac{u\, du}{u^2 - u + 1}$.

120. $\int u e^{u^2}\, du$.

121. $\int \frac{du}{u(\ln u)^n}$.

122. $\int \frac{8x}{3 + 4x}\, dx$.

123. $\int \frac{dx}{x^2 + x + 1}$.

124. $\int \frac{dx}{x^2 + 2ax + b}$.

125. $\int t^2 \sqrt{1 + t^3}\, dt$.

126. $\int \frac{t + 1}{\sqrt{1 - t^2}}\, dt$.

127. $\int \frac{t^4}{1 - t}\, dt$.

128. $\int \cos^n t \sin t\, dt$.

129. Докажите, что

$$\int \frac{dx}{a^2 - x^2} = \frac{1}{a} \operatorname{Arth} \frac{x}{a}, \qquad \int \frac{dx}{\sqrt{a^2 - x^2}} = \operatorname{Arsh} \frac{x}{a}.$$

(Сравните с примерами ж), з).)

Б) Правило дифференцирования произведения (стр. 451)

$$(p(x)\cdot q(x))' = p(x)\cdot q'(x) + p'(x)\cdot q(x)$$

в интегральной форме записывается следующим образом:

$$p(x)\cdot q(x) = \int p(x)\cdot q'(x)\,dx + \int p'(x)\cdot q(x)\,dx,$$

или же

$$\int p(x)\cdot q'(x)\,dx = p(x)\,q(x) - \int p'(x)\cdot q(x)\,dx. \qquad \text{(II)}$$

В этой форме оно называется *правилом интегрирования по частям.* Это правило бывает полезно в тех случаях, когда функция, стоящая под интегралом, имеет вид $p(x)\,q'(x)$, причем неопределенный интеграл $q(x)$ от функции $q'(x)$ известен. Формула (II) сводит проблему неопределенного интегрирования функции $p(x)\,q'(x)$ к проблеме интегрирования функции $p'(x)\,q(x)$, что часто оказывается более простым.

а) $J = \int \ln x\, dx$. Положим $p(x) = \ln x$, $q'(x) = 1$, так что $q(x) = x$. Тогда формула (II) нам дает

$$\int \ln x\, dx = x\ln x - \int \frac{x}{x}\,dx = x\ln x - x.$$

б) $J = \int x\ln x\, dx$. Положим $p(x) = \ln x$, $q'(x) = x$. Тогда

$$J = \frac{x^2}{2}\ln x - \int \frac{x^2}{2x}\,dx = \frac{x^2}{2}\ln x - \frac{x^2}{4}.$$

в) $J = \int x\sin x\, dx$. На этот раз положим $p(x) = x$, $q(x) = -\cos x$ и получим

$$\int x\sin x\, dx = -x\cos x + \sin x.$$

Вычислите по частям следующие интегралы:

130. $\int xe^x\,dx$.

131. $\int x^2\cos x\,dx$.

(*Указание*: примените (II) дважды.)

132. $\int x^a\ln x\,dx$ $(a \neq -1)$.

133. $\int x^2e^x\,dx$.

(*Указание*: воспользуйтесь упражнением 130.)

Интегрируя по частям $\int \sin^m x\,dx$, мы получаем замечательную формулу для числа π в виде бесконечного произведения. Напишем функцию $\sin^m x$ в виде $\sin^{m-1} x\cdot\sin x$ и проинтегрируем по частям в пределах

от 0 до $\frac{\pi}{2}$. Тогда получим

$$\int_0^{\frac{\pi}{2}} \sin^m x\,dx = (m-1)\int_0^{\frac{\pi}{2}} \sin^{m-2} x \cos^3 x\,dx =$$

$$= -(m-1)\int_0^{\frac{\pi}{2}} \sin^m x\,dx + (m-1)\int_0^{\frac{\pi}{2}} \sin^{m-2} x\,dx,$$

или же

$$\int_0^{\frac{\pi}{2}} \sin^m x\,dx = \frac{m-1}{m}\int_0^{\frac{\pi}{2}} \sin^{m-2} x\,dx$$

$\Big($так как первый член в правой части (II), pq, обращается в нуль при $x = 0$ и $x = \frac{\pi}{2}\Big)$. Применяя повторно последнюю формулу, найдем следующие значения интегралов $I_m = \int_0^{\frac{\pi}{2}} \sin^m x\,dx$ (формулы различаются в зависимости от четности n):

$$I_{2n} = \frac{2n-1}{2n}\cdot\frac{2n-3}{2n-2}\cdot\ldots\cdot\frac{1}{2}\cdot\frac{\pi}{2},$$

$$I_{2n+1} = \frac{2n}{2n+1}\cdot\frac{2n-2}{2n-1}\cdot\ldots\cdot\frac{2}{3}.$$

Так как $0 < \sin x < 1$ при $0 < x < \frac{\pi}{2}$, то $\sin^{2n-1} x > \sin^{2n} x > \sin^{2n+1} x$, и следовательно,

$$I_{2n-1} > I_{2n} > I_{2n+1}$$

(см. стр. 437), или

$$\frac{I_{2n-1}}{I_{2n+1}} > \frac{I_{2n}}{I_{2n+1}} > 1.$$

Подставляя в эти неравенства вычисленные значения интегралов, мы получаем

$$\frac{2n+1}{2n} > \frac{1\cdot3\cdot3\cdot5\cdot5\cdot7\cdot\ldots\cdot(2n-1)(2n-1)(2n+1)(2n+1)}{2\cdot2\cdot4\cdot4\cdot6\cdot\ldots\cdot(2n)(2n)}\cdot\frac{\pi}{2} > 1.$$

Остается положить $n \to \infty$; тогда, убедившись, что средняя часть неравенства стремится к 1, мы получаем следующее принадлежащее Уоллису представление для числа $\frac{\pi}{2}$:

$$\frac{\pi}{2} = \frac{2\cdot2\cdot4\cdot4\cdot6\cdot6\cdot\ldots\cdot2n\cdot2n\ldots}{1\cdot3\cdot3\cdot5\cdot5\cdot7\cdot\ldots\cdot(2n-1)(2n-1)(2n+1)\ldots} =$$

$$= \lim \frac{2^{4n}(n!)^4}{[(2n)!]^2(2n+1)} \quad \text{при} \quad n\to\infty.$$

ДОБАВЛЕНИЕ 1

Вклейка «От издательства» в первое издание книги на русском языке*

Существует большой разрыв между математикой, которая преподается в средней школе, и наиболее живыми и важными для естествознания и техники разделами современной математической науки. Наиболее существенной стороной этого разрыва является отсутствие в курсе средней школы элементов математического анализа, которые совершенно необходимы для понимания основных идей физики и многих разделов техники.

Курсы высшей математики для техников, химиков, биологов и специалистов по сельскому хозяйству в наших вузах содержат достаточно солидное изложение элементов классического анализа, но оставляют совершенно в стороне ряд более общих и новых идей математики, относящихся, например, к проективной геометрии, топологии, более высоким разделам вариационного исчисления и т. п. Между тем, эти идеи становятся все более существенными для всей совокупности точных и технических наук.

Наконец, молодежь, избирающая своей специальностью математику или те разделы естественных наук (механику, астрономию, физику), изучение которых в высшей школе связано с прохождением вполне современного большого курса математики, часто нуждается в том, чтобы еще на стадии перехода из средней школы в высшую в более легкой и наглядной форме познакомиться с различными разделами математики вплоть до наиболее высоких и современных.

Выпускаемая в русском переводе книга Р. Куранта и Г. Роббинса может в некоторой мере заполнить указанные выше разрывы между систематическими учебными курсами математики и естественными запросами различных категорий читателей в направлении общего ознакомления с более высокими разделами математики.

Отдельные главы этой книги в значительной мере независимы друг от друга (см. указания автора «Как пользоваться книгой») и могут пред-

---

* О причинах, вызвавших появление этой вклейки, см. предисловие В. М. Тихомирова, с. 5. — *Прим. ред. наст. изд.*

ставить интерес в первую очередь для следующих категорий читателей.

1. Главы VI–VIII позволяют читателям с подготовкой в размере курса средней школы в сравнительно легкой форме познакомиться с основными идеями высшей математики (дифференциальное и интегральное исчисления, дифференциальные уравнения, вариационное исчисление).

2. Читателям, прошедшим краткий курс высшей математики, — инженерам, химикам, многим преподавателям математики в средней школе — будут по преимуществу интересны главы I–V, вводящие их в менее знакомые им разделы математики, и некоторый дополнительный материал в последних главах. Главы I–V будут интересны также тем выпускникам средней школы, которые пожелают, в связи с выбором специальности, познакомиться с современной математикой.

3. Наконец, многие разделы книги могут быть использованы в школьных кружках любителей математики и в кружках и семинарах для студентов младших курсов физико-математических факультетов университетов, педагогических и учительских институтов.

Наша отечественная литература, обслуживающая перечисленные потребности, еще недостаточна. Поэтому Издательству представлялось весьма желательным ее пополнение хорошо написанной переводной книгой, хотя бы и содержащей некоторые недостатки и ошибки.

Первый из авторов книги, ответственный за ее общий замысел,— Р. Курант является крупным математиком, имеющим заслуги по преимуществу в областях математического анализа, близких к вопросам математической физики. §§ 7–11 отражают в элементарных рамках данной книги собственные исследования Р. Куранта.

Теория чисел, геометрия и топология более далеки от личных интересов Р. Куранта. Выбор вопросов из этих областей иногда имеет несколько случайный характер, а их изложение в отдельных случаях не вполне точно. Однако подход к этим вопросам с точки зрения математика, привыкшего работать в области математического естествознания, иногда интересен и своеобразен.

Не ставя своей специальной целью излагать историю идей и методов математики, Курант, однако, не может избежать замечаний исторического характера. Последние крайне немногочисленны и явно неполны. Так, например, отмечая выдающихся математиков, внесших вклад в теорию чисел, Курант совершенно не упоминает великого русского аналитика П. Л. Чебышёва; говоря о развитии современной топологии, проходит мимо достижений школы советских топологов. В исторических ссылках Куранта имеются и прямые ошибки. Приоритет открытия неевклидовой геометрии бесспорно принадлежит великому русскому геометру Н. И. Лобачевскому; Курант этого не подчеркивает, и у читателя создается впечатление, что автор отдает предпочтение в этом вопросе Гауссу, который, владея лишь замыслом неевклидовой геометрии, не только не

дал этому замыслу достаточного развития, не только не опубликовал своих взглядов на этот вопрос, но и не позволял опубликовывать их тем, кому они были известны. Неправильно распределяет Курант заслуги между Зигелем и Гельфондом: решение общей проблемы Гильберта (доказательство трансцендентности чисел вида α^β, где α — алгебраическое, а β — алгебраическое иррациональное число) целиком принадлежит А. О. Гельфонду. Тенденциозно и умаление заслуг И. М. Виноградова.

Наконец, в принципиальных философских вопросах математики Р. Курант является эклектиком. Поэтому Издательство предпочло сократить авторское введение «Что такое математика». Советскому читателю излишне пояснять, что пожелания автора относительно будущего математики, которые заканчивают это введение, не могут быть осуществлены буржуазной наукой. Это — задача советской математики.

ДОБАВЛЕНИЕ 2

О создании книги «Что такое математика?»*

Со времени своего приезда в США в 1934 г. Курант обдумывал научные нужды своей новой родины. Он рассматривал их гораздо шире, чем возможности одного отдельно взятого университета. В США надо было создать национальный научный центр — подобный École Polytechnique во Франции — который выпускал бы хорошо подготовленную научную элиту, достаточно подготовленную для работы в условиях надвигающейся войны и последующих трудных лет.

Первые два варианта заметки «О национальном Институте для изучения фундаментальных и прикладных наук» не имеют даты, а третий вариант помечен зимой 1940–41 гг.

С первых же абзацев этого текста слышен голос Феликса Клейна, который восхищался программой и подходом к образованию в École Polytechnique и всегда жалел, что ее идеалы (тесная связь чистой и прикладной науки, сочетание учебы и исследований, личный контакт преподавателя со студентами) «никогда не имели крепких корней на германской почве»...

Курант в своей записке подчеркивает своеобразие исторической обстановки, в которой во Франции родилась École Polytechnique. После революции Франция была «экономически разорена, интеллектуально и морально неустойчива».

После войны со всей Европой образовательные учреждения были дезорганизованы. Ученые, «понимавшие ситуацию и обладавшие инициативой», выработали план института высшего образования «на чрезвычайно высоком, по сравнению с прежним, уровне». Предполагалось, что студенты будут демократично, но тщательно отбираться, а преподавателями станут лучшие ученые страны. Новое учреждение вскоре оправдало «высочайшие надежды» своих основателей. Менее чем через два года армия, флот, промышленность и правительство уже стали получать в свои ряды людей, чье образование было лучшим в мире...

В 1940 г. Курант несколько раз переделывал свою заметку — правда,

---

* Воспроизводится по книге: Reid C. Courant in Göttingen and New York. The Story of an Improbable Mathematician. — New York: Springer-Verlag, 1976. Перевод *Е. А. Коноваленко*. — *Прим. ред. наст. изд.*

изменения касались скорее слов, чем содержания,— наконец, в начале 1941 года он решил (став к тому времени гражданином США) обнародовать свои предложения. Он считал, что планируемый им Институт может начать работать довольно скоро, если его начать пока с курсов математики и физики. В конце записки, датированной зимой 1940–41 гг., он оптимистично указал время открытия Института — сентябрь 41 г.

В продолжение учебного 1940–41 учебного года Курант занимался еще одним проектом, который он также считал частью патриотического служения своей новой родине,— книгой «What is Mathematics?». Он работал над ней уже почти 5 лет и привлек к этой работе некоторых студентов.

Давид Гильбарг писал конспекты его лекций, а еще семеро молодых людей, включая сына Куранта, Эрнста, помогали (как сказано в предисловии) «в бесконечной работе по написанию и переписыванию этого труда».

Весной 1939 г. Курант решил, что предмет книги будет слишком узок, если ограничиваться только его собственными интересами. Во время поездки в Принстон он советовался с разными людьми, и Марстон Морс порекомендовал ему в помощники своего ассистента Герберта Роббинса, молодого тополога из Гарварда. Курант встретился с Роббинсом в его офисе.

Когда я [*К. Рид*] разговаривала с Роббинсом в 1975 г. в его квартире возле Колумбийского Университета, он не помнил уже хорошо, в первую ли встречу, или в одну из следующих, Курант предложил ему работать над книгой «What is Mathematics?». Роббинс рассказывал, что приехав в Нью-Йоркский университет в конце 1939 г., он преподавал там элементарные предметы днем и читал более сложные лекции по вечерам. Курант передал ему все, что уже сделали его прежние помощники, поговорил о концепции книги в целом; и попросил прочесть весь труд, улучшив и дополнив его.

Я спросила Роббинса, как они с Курантом работали над книгой. «Трудно сказать, — отвечал Роббинс.— У него были мимеографические записки одного курса лекций, который он читал когда-то прежде, эти лекции были записаны кем-то из студентов — и это составляло примерно четверть или треть того материала, который в конечном итоге вошел в книгу. Некоторые главы были там в окончательной форме, других не было вовсе.

Например, одну из глав книги мы хотели посвятить топологии, и обсуждали, что в этой главе должно быть. О некоторых вещах у него были очень четкие понятия, о некоторых у меня. Два года я работал и показывал ему, что получилось — он комментировал и критиковал, и я переделывал заново... Иногда он придумывал интересные решения, иногда я... Я бы не сказал, что было что-то особенное в способе нашей

совместной работы. Это было довольно тесное сотрудничество, хотя мы никогда не садились писать вместе.

Роббинс говорил, что не слушал ни этого курса лекций Куранта, ни какого-либо другого.

Сперва, по словам Куранта, от молодого Роббинса было немного помощи. «Он даже был мне помехой, так как работал не очень-то много. Но потом, после доверительного разговора между ними, Роббинс был весьма полезен». В конце концов, как вспоминает Роббинс, Курант сказал ему, что очень доволен его работой. Деньги из фонда Рокфеллера ($1500), из которых Курант выплачивал Роббинсу за работу, подошли к концу. (Кроме этого Роббинс получал $2500 в год как преподаватель университета.) По словам Роббинса, Курант предложил ему быть соавтором, так как их сотрудничество оказалось более продуктивным, чем это задумывалось в начале работы...

Роббинс рассказывал, что поначалу он принял участие в этой работе главным образом потому, что хотел подзаработать. «Сперва мне не очень нравилось заниматься этой книгой, потому что это отнимало довольно много времени, а вы понимаете, что молодому человеку, только что получившему степень (Ph. D), для создания научной репутации необходимо больше заниматься исследованиями, нежели популяризаторством. Так что я колебался, стоит ли тратить еще года полтора на то, что я считал отвлечением от занятий, которые меня действительно интересовали... Я не ожидал, что стану соавтором. Но когда он предложил мне это, я согласился. Я уже был довольно сильно увлечен этим делом к тому времени»...

Проскауэр посоветовал Куранту обратиться по поводу издания книги не к столь специальному издательству, каким было InterScience, — чтобы у книги был более широкий круг читателей. И в начале 1941 г. Курант провел переговоры с из издательством Macgrow-Hill, которое ранее уже проявляло интерес к книге «What is Mathematics?». Ему пришло в голову, что эта книга должна послужить еще и своего рода зацепкой — чтобы заинтересовать большое, солидное американское издательство и в некоторых других имевшихся у Куранта идеях. Едва подписав контракт, Курант начал набрасывать план серии учебников по математике — по образцу его серии в издательстве Springer.

Во время совместной с Роббинсом работы над книгой Курант заботился и о научном будущем своего молодого соавтора... Он поручил Роббинсу читать в университете курс лекций по вероятности и статистике.

«Я узнал об этом всего за несколько недель до начала лекций, — вспоминает Роббинс. — А до этого у меня не было ни интереса, ни даже слабого знакомства ни с теорией вероятностей, ни со статистикой.»

Работа Роббинса над этим курсом произвела на Куранта большое

впечатление. Он считал, что молодому человеку было бы желательно изучить статистику и теорию вероятностей «из источника». Источником, по мнению Куранта, был Джерси Нейман, знаменитый польский ученый в этой области, который как раз недавно приехал в Беркли. Ранней весной 1941 г. Курант обратился с письмом к Гриффиту Эвансу, возглавлявшему математический факультет в Калифорнийском университете. Предложение Куранта состояло в том, чтобы оказать Роббинсу финансовую помощь (тот был довольно беден, да к тому же содержал мать и младшую сестру) для занятий с Нейманом в течение лета...

Теперь Роббинс — один из выдающихся специалистов в теории вероятностей и статистики, и я спросила его, действительно ли у него была возможность поехать летом 41 г. в Беркли для занятий с Нейманом. «Нет,— сказал он. Он даже не знал, что Курант предлагал это Эвансу. — Если бы это произошло, уверен, моя жизнь могла сложиться иначе, ведь мне удалось встретиться с Нейманом намного позже».

Всю весну 1941 г. Курант был очень занят. Он старался закончить свою книгу, заинтересовать людей в своей идее национального научного института, а также пытался ввести в Нью-Йоркском университете ряд летних математических курсов, ориентированных на нужды армии. Поэтому он был очень расстроен, когда люди из Macgrow-Hill выразили сомнения по поводу коммерческой выгоды от издания «What is Mathematics?», хотя и настаивали на своей готовности ее издать. Он решил, что книга должна быть в печати к концу 1941 г., и не хотел идти ни на какие уступки касающиеся этой книги.

У Куранта уже был опыт издания книг со своим участием в прибыли. Теперь он решил сам стать своим издателем. Летом 41 г. он занял денег у обеспеченных друзей и договорился с Waverly-Press о напечатании своей книги. Затем он подписал контракт с Oxford-Press о распространении книги...

Летом 1941 г. большая часть книги была уже в печати, и Курант собрался написать предисловие. В 30-е годы на английском языке уже вышло в свет несколько популярных книг по математике. Но Курант чувствовал, что у него получилась книга, совсем не похожая на них. По мнению Куранта, у всех них был серьезный недостаток. Они были написаны, исходя из неверных позиций. Понимания математики нельзя достичь путем легких развлечений, равно как никакое, даже самое блестящее описание, не сможет передать понимание музыки тому, кто никогда внимательно в музыку не вслушивался. Чтобы понять математику, надо ею заниматься. И Курант в своей книге хотел дать читателю возможность «действительно прикоснуться к содержанию живой математики»...

Некоторую озабоченность у Куранта вызывало название книги. Оно казалось ему «слегка нечестным». Однажды на вечеринке у Г. Вейля

он спросил совета у Томаса Манна. Должна ли книга называться «Что такое математика?», или ее следует назвать примерно так: «Математические дискуссии по поводу основных элементарных задач для широкой публики»,— это название более точно соответствует содержанию, но «немного скучнее». «Манн сказал, что не может дать мне совета, но может поделиться собственным опытом», — вспоминал Курант. — «Среди его книг, переведенных на английский, была „Лотта в Веймаре“. Незадолго до выхода этой книги в свет к нему пришел м-р Кнопф и сказал: „Теперь нам надо выбрать название; вот моя жена, которая знает толк в этих вещах, считает, что можно озаглавить книгу «Возвращение любимого»“. Манну это не слишком понравилось: в конце концов, „Лотта в Веймаре“ одинаково хорошо звучит, что по-английски, что по-немецки. Но м-р Кнопф сказал, что у него есть одно замечание: „Если мы напечатаем «Лотту в Веймаре», то мы сможем продать 10 или 20 тысяч экземпляров, а если мы напечатаем «Возвращение любимого», то можно продать и 100 тысяч — и авторский гонорар будет соответствующим.“ Манн сказал, что он согласен, пусть будет „Возвращение любимого“.» Курант поблагодарил Манна — и позвонил в издательство...

Хотя Роббинс читал гранки книги и несколько раз ездил в издательство в Балтимор, он еще не видел титульного листа. И вот, в августе 41 г., он увидел его: «Рихард Курант. Что такое математика?»

«Когда я это первый раз увидел, я вдруг сказал: „Боже мой, этот человек — обманщик“. Это было вроде холодного душа. В тот момент я пожалел не столько о том, что не увижу своего имени на обложке книги (потому что сразу решил, что все-таки увижу), сколько о том, что это был конец моего отношения к Куранту как к приличному, достойному человеку, который хотел способствовать труду своего молодого коллеги, не заботясь о собственном престиже, etc... Позднее, конечно, мне приходилось не раз слышать о нем: „Грязный Дик“. Люди не удивлялись, что такое могло случиться, потому что слышали и другие подобные истории. Я же тогда еще ничего не слышал, и знал о Куранте только хорошее... Я даже любил его.»

Роббинс посоветовался с некоторыми людьми в Washington Square, что ему делать. «Они говорили: „Ну, ты понимаешь, в Германии такое случается довольно часто. Многие книги знаменитых профессоров на самом деле написаны кем-то из их студентов, в качестве части общения“. Я сказал, что, во-первых, я не студент Куранта; во-вторых, здесь не Германия; и в-третьих, мне это просто не нравится.»

Теперь уже невозможно установить точную последовательность событий тех дней, не нашедших отражения в переписке Куранта и Роббинса. Но в некоторый момент, как рассказывал мне Роббинс, он поговорил с Хаслером Уитни, под чьим руководством он работал в Гарварде.

«Когда я обо всем рассказал Уитни, это привело его в большое него-

дование, и он сказал: „Хорошо, скажите Куранту, что если он будет продолжать в том же духе, то на следующем заседании Американского Математического общества я подниму этот вопрос — и мы исключим его из членов Математического общества“.»

Как вспоминает Роббинс, он «пообещал или даже пригрозил» в письме Куранту от 17 августа 41 г. высказать все свои чувства и мысли по поводу того, что было написано на обложке книги. Он отложил это на некоторое время, потому что его переживания по этому поводу были слишком сильны, и он не надеялся на спокойную беседу с Курантом, а «жаркий спор мог иметь плохие последствия для самой книги и для даты ее выхода в свет».

В своем письме от 17 августа 41 г. Роббинс утверждал, что хотя он понимает, что эта книга является, в основном, детищем Куранта, все же и сам он отдал ей так много сил и так сильно был эмоционально вовлечен в написание книги, что увидеть свое имя на обложке рядом с именем Куранта — очень важно для него. Кроме того, хотя в Европе другие обычаи, в Америке принято решать эти вопросы именно так: все признают, что первое имя на обложке — это имя настоящего автора книги; но указывают и второе имя — его молодого помощника, коллеги. Что касается до финансовых вопросов, он (Роббинс) предпочитает оставить их целиком на усмотрение Куранта. Он только просит, чтобы на обложке было написано «Р. Курант и Г. Роббинс».

В письме Роббинса не упомянуты ни угроза Уитни, ни намерения Роббинса действовать в защиту своих прав. Роббинс считает, что Курант прослышал об этом от других людей. Как бы то ни было, после получения этого письма Курант согласился изменить титульный лист. Роббинс объяснял мне, что он старался написать это письмо так, чтобы сделать идею соавторства более приемлемой для Куранта. Ему это удалось. Когда Курант показывал мне это письмо, он сказал, что, по его ощущениям, оно очень точно отражает ситуацию, возникшую в то время между ним и Роббинсом.

В течение следующих нескольких недель осенью 41 г. Курант постепенно пришел к осознанию того, что роль Роббинса в написании книги действительно заслуживает названия соавторства. И 28 сентября 41 г. он написал молодому человеку длинное строгое письмо. «У меня создалось впечатление, что Вы позволили себе занять (или кто-то Вас к этому подтолкнул) очень неловкую психологическую позицию. Я думаю, что Вам необходимо отдавать себе отчет в том, как в действительности обстоят дела. Дело не в том, сколько сил, времени, энергии, усердия Вы потратили на эту работу. Эта книга — мое детище по замыслу, по планированию, по содержанию, по ее математическим идеям, — и она выражает именно мои личные взгляды и цели в большей степени, чем любая другая из моих публикаций. Вы достаточно индивидуальны, чтобы иметь свои

собственные взгляды, которые вовсе не обязаны совпадать с моими, и было бы вполне естественно, чтобы Ваши взгляды заметно отличались от моих,... чтобы развиться в будущем. По этой причине я был и остаюсь озабочен тем, чтобы вопрос об авторстве ясно понимался всеми, и в первую очередь Вами. Дело не в амбициях, а в сущности научной ответственности. Конечно, я нуждался в помощнике... Ваш вклад в эту работу далеко превзошел все, что я мог ожидать от грамотного математика. Я нисколько не хочу лишать Вас похвалы и общественного признания. И когда Вы, в письме от 17 августа, настаивали на том, чтобы обеспечить Вам такое признание, поместив Ваше имя на обложку, я немедленно согласился. Ваше письмо снова уверило меня в том, что между нами не было и никогда не могло быть существенного непонимания по вопросам обоснования авторства, и что ни у кого не было намерений произвести на публику неверное впечатление по этому поводу... »

Книга Рихарда Куранта и Герберта Роббинса «Что такое математика?» имела гораздо больший успех, чем кто бы то ни было (кроме, возможно, Куранта) мог предполагать. Со времени своего выхода в свет она была переведена на несколько языков и разошлась более чем в 100 000 экземплярах. Ее часто называют «математическим бестселлером»...

Сам же Курант, несмотря на весь успех, был «слегка разочарован» в этой книге. Успех все же не достиг того уровня, чтобы заметно повлиять на широкий круг «образованных любителей», которых Курант надеялся приобщить к некоторым красотам математики.

Рекомендуемая литература

С момента выхода последнего издания книги «Что такое математика?» прошло много лет. С тех пор по элементарной математике и ее связям с современной наукой вышло множество изданий. Ниже мы перечисляем некоторые книги, посвященные этой тематике. Они предназначены для широкого круга читателей: от школьников 6–7 классов до студентов 1–2 курсов и преподавателей математики. (Мы не включили в этот список различные сборники задач.)

* * *

Прежде всего, нам хочется порекомендовать для чтения журнал «Квант», где опубликовано огромное количество статей по самым разным вопросам.

Отметим также ряд книжных серий (как правило, мы не включили книги из этих серий в наш указатель), доступных по своему уровню школьникам: «Популярные лекции по математике», «Библиотека математического кружка», «Библиотечка „Квант“» (издательство «Наука»); «Современная математика. Вводные курсы» (издательство «Мир»).

Следует также отметить как ранние выпуски журнала «Математическое просвещение», так и возобновившиеся с 1997 г. выпуски, публикуемые Московским Центром непрерывного математического образования (к настоящему моменту вышло 10 выпусков).

Для любителей истории науки рекомендуем сборники «Историко-математические исследования», публикуемые сектором математики Института истории естествознания и техники.

Ниже мы приводим книги по различным разделам математики. Однако при этом не следует забывать о единстве математики и помнить об условности любых перегородок в науке.

Общие вопросы математики

[1] Адамар Ж. Психология процесса изобретения в области математики. — М.: МЦНМО, 2001.

[2] Вейль Г. Симметрия. — М.: Наука, 1968.

[3] Вейль Г. Математическое мышление. — М.: Наука, 1990.

[4] Вилейтнер Г. История математики от Декарта до середины XIX столетия. — М.: Физматгиз, 1966.

[5] Варга Б., Димень Ю., Лопариц Э. Язык, музыка, математика. — М.: Мир, 1981.

[6] Гиндикин С. Г. Рассказы о физиках и математиках. — М.: МЦНМО, 2001.

[7] Кириллов А. А. Что такое число? — М.: Наука, 1993.

[8] Клайн М. Математика. Утрата определенности. — М.: Мир, 1984.

[9] Клейн Ф. Лекции о развитии математики в XIX столетии. — М.: Наука, 1989.

[10] Клейн Ф. Элементарная математика с точки зрения высшей. Т. 1, 2. — М.: Наука, 1987.

[11] Клини С. К. Введение в метаматематику. — М.: ИЛ, 1957.

[12] Колмогоров А. Н. Математика — наука и профессия. — М.: Наука, 1987.

[13] Кымпан Ф. История числа π. — М.: Наука, 1971.

[14] Литлвуд Дж. Математическая смесь. — М.: Наука, 1990.

[15] Марков А. А. Конструктивное направление (в математике и логике) // Философская энциклопедия, т. 3, М., 1964.

[16] Пидоу Д. Геометрия и искусство. — М.: Мир, 1979.

[17] Пойа Д. Математика и правдоподобные рассуждения. — М.: Наука, 1975.

[18] Пойа Д. Математическое открытие. Решение задач: основные понятия, изучение и преподавание. — М.: Наука, 1976.

[19] Прасолов В. В. Рассказы о числах, многочленах и фигурах. — М.: Фазис, 1997

[20] Пуанкаре А. О науке. — М.: Наука, 1990.

[21] Радемахер Г., Теплиц О. Числа и фигуры. Опыт математического мышления. — М.: Физматгиз, 1966.

[22] Рассказы о математике и математиках. — М.: МЦНМО, 2000.

[23] Розенфельд Б. А. История неевклидовой геометрии. — М.: Наука, 1976.

[24] Сингх С. Великая теорема Ферма. История загадки, которая занимала лучшие умы мира на протяжении 358 лет. — М.: МЦНМО, 2000.

[25] Сойер У. У. Прелюдия к математике. — М.: Просвещение, 1972.

[26] Стройк Д. Я. Краткий очерк истории математики. — М.: Наука, 1984.

[27] Фрейденталь Г. Математика в науке и вокруг нас. — М.: Мир, 1977.

[28] Энциклопедия элементарной математики. Т. 1–5. Под ред. П. С. Александрова, А. И. Маркушевича и А. Я. Хинчина. — М.: ГТТИ, 1952–1966.

Принцип математической индукции, теория множеств, математическая логика

[29] Верещагин Н. К., Шень А. Лекции по математической логике и теории алгоритмов: в 3-х ч. М.: МЦНМО. Ч. 1. Начала теории множеств, 1999; Ч. 2. Языки и исчисления, 2000; Ч. 3. Вычислимые функции, 1999.

[30] Виленкин Н. Я. Комбинаторика. — М.: Наука, 1969.

[31] Виленкин Н. Я. Рассказы о множествах. — М.: Наука, 1969.

[32] Гжегорчик А. Популярная логика. Общедоступный очерк логики предложений. — М.: Наука, 1979.

[33] Градштейн И. С. Прямая и обратная теоремы. — М.: Наука, 1973.

[34] Кац М., Улам С. Математика и логика. Ретроспектива и перспектива. — М.: Мир, 1971.

[35] Манин Ю. И. Вычислимое и невычислимое. — М.: Сов. радио, 1980.

[36] Манин Ю. И. Доказуемое и недоказуемое. — М.: Сов. радио, 1978.

[37] Мендельсон Э. Введение в математическую логику. — М.: Наука, 1984.

[38] Новиков П. С. Элементы математической логики. М.: Физматгиз, 1959.

[39] Соминский И. С., Головина Л. И., Яглом И. М. О математической индукции. — М.: Наука, 1977.

[40] Успенский В. А. Треугольник Паскаля. — М.: Наука, 1979.

[41] Успенский В. А. Теорема Гёделя о неполноте. — М.: Наука, 1982.

[42] Успенский В. А. Машина Поста. — М.: Наука, 1988.

[43] Успенский В. А. Что такое нестандартный анализ. — М.: Наука, 1987.

Алгебра и теория чисел

[44] Айерланд К., Роузен М. Классическое введение в современную теорию чисел. — М.: Мир, 1987.

[45] Александров П. С. Введение в теорию групп. — М.: Наука, 1980.

[46] Алексеев В. Б. Теорема Абеля в задачах и решениях. — М.: МЦНМО, 2001.

[47] Аршинов М. Н., Садовский Л. Е. Коды и математика. Рассказы о кодировании. — М.: Наука, 1983.

[48] Берман Г. Н. Число и наука о нем. — М.: ГТТИ, 1949.

[49] Болтянский В. Г., Виленкин Н. Я. Симметрия в алгебре. — М.: Наука, 1967.

[50] Винберг Э. Б. Курс алгебры. — М.: Факториал Пресс, 2001.

[51] Коблиц Н. p-адические функции, p-адический анализ и дзета-функции. М.: Мир, 1982.

[52] Курош А. Г. Курс высшей алгебры. — М.: Наука, 1975.

[53] Литцман В. Великаны и карлики в мире чисел. — М.: Физматгиз, 1959.

[54] Понтрягин Л. С. Знакомство с высшей математикой: Алгебра. — М.: Наука, 1987.

[55] Постников М. М. Теория Галуа. — М.: Физматгиз, 1963.

[56] Прасолов В. В. Многочлены. — М.: МЦНМО, 2000.

[56] Прасолов В. В., Соловьев Ю. П. Эллиптические функции и алгебраические уравнения. — М.: Факториал, 1997.

[57] Проскуряков И. В. Числа и многочлены. — М.: Просвещение, 1965.

[58] Рид М. Алгебраическая геометрия для всех. — М.: Мир, 1991.

[59] Серпинский В. 250 задач по элементарной теории чисел. — М.: Просвещение, 1968.

[60] Солодовников А. С. Ведение в линейную алгебру и линейное программирование. — М.: Просвещение, 1966.

[61] Трост Э. Простые числа. — М.: Физматгиз, 1959.

[62] Фрид Э. Элементарное введение в абстрактную алгебру. — М.: Мир, 1979.

[63] Хассе Г. Лекции по теории чисел. — М.: ИЛ, 1953.

[64] Шафаревич И. Р. Основные понятия алгебры. — Итоги науки и тех-

ники. Современные проблемы математики. Фундаментальные направления. Т. 11. М.: ВИНИТИ, 1986.

[65] Эдвардс Г. Последняя теорема Ферма. Генетическое введение в алгебраическую теорию чисел. — М.: Мир, 1980.

Геометрия

[66] Адамар Ж. Элементарная геометрия. Т. 1, 2. — М.: Учпедгиз, 1948–51.

[67] Александров П. С. Курс аналитической геометрии и линейной алгебры. — М.: Наука, 1979.

[68] Берже М. Геометрия. Т. 1, 2. — М.: Мир, 1984.

[69] Берман Г. Н. Циклоида. Об одной замечательной кривой линии и некоторых других, с ней связанных. — М.: Наука, 1980.

[70] Балк М. Б., Болтянский В. Г. Геометрия масс. — М.: Наука, 1987.

[71] Болтянский В. Г., Ефремович В. А. Наглядная топология. — М.: Наука, 1982.

[72] Борисович Ю. Г., Близняков Н. М., Израилевич Я. А. и др. Введение в топологию. — М.: Физматлит, 1995.

[73] Васильев Н. Б., Гутенмахер В. Л. Прямые и кривые. — М.: МЦНМО, 2000.

[74] Веннинджер М. Модели многогранников. — М.: Мир, 1974.

[75] Гальперин Г. А., Земляков А. Н. Математические биллиарды. — М.: Наука, 1990.

[76] Гельфанд И. М. Лекции по линейной алгебре. — М.: Добросвет, МЦНМО, 1998.

[77] Гильберт Д., Кон-Фоссен С. Наглядная геометрия. — М.: Наука, 1981.

[78] Глаголев Н. А. Проективная геометрия. — М.–Л.: ОНТИ, 1936.

[79] Кокстер Х. С. М. Введение в геометрию. — М.: Наука, 1966.

[80] Кокстер Г. С., Грейтцер С. Л. Новые встречи с геометрией. — М.: Мир, 1978.

[81] Косневски Ч. Начальный курс алгебраической топологии. — М.: Мир, 1983.

[82] Кроуэлл Р., Фокс Р. Введение в теорию узлов. — М.: Мир, 1967.

[83] Клейн Ф. Неевклидова геометрия. — М.–Л.: ОНТИ, 1936.

[84] Литцман В. Старое и новое о круге. — М.: Физматгиз, 1960.

[85] Литцман В. Теорема Пифагора. — М.: Физматгиз, 1960.

[86] Никулин В. В., Шафаревич И. Р. Геометрии и группы. — М.: Наука, 1983.

[87] Милнор Дж., Уоллес А. Дифференциальная топология. Начальные главы. — М.: Мир, 1972.

[88] Прасолов В. В., Тихомиров В. М. Геометрия. — М.: МЦНМО, 1997.

[89] Прасолов В. В. Геометрические задачи Древнего мира. — М.: Фазис, 1997.

[90] Прасолов В. В. Задачи по планиметрии. — М.: Наука, 1995.

[91] Прасолов В. В. Геометрия Лобачевского. — М.: МЦНМО, 2000.

[92] Торп Дж. Начальные главы дифференциальной геометрии. — М.: Мир, 1982.

[93] Тужилин А. А., Фоменко А. Т. Элементы геометрии и топологии минимальных поверхностей. — М.: Наука, 1991.

[94] Уокер Р. Алгебраические кривые. — М.: ИЛ, 1952.

[95] Яглом И. М. Принцип относительности Галилея и неевклидова геометрия. — М.: Наука, 1969.

Математический анализ

[96] Арнольд В. И. Гюйгенс и Барроу, Ньютон и Гук — первые шаги математического анализа и теории катастроф от эвольвент до квазикристаллов. 1989

[97] Берс Л. Математический анализ. Т. 1, 2. — М.: Высшая школа, 1975.

[98] Гельфанд И. М., Глаголева Е. Г., Кириллов А. А. Метод координат. — М.: Наука, 1973.

[99] Гельфанд И. М., Глаголева Е. Г., Шноль Э. Э. Функции и графики. — М.: МЦНМО, 2001.

[100] Зельдович Я. Б. Высшая математика для начинающих и ее приложения к физике. — М.: Наука, 1970.

[101] Зорич В. А. Математический анализ.: в 2-х т. — М.: МЦНМО. Т. 1. 2001; Т. 2. 1998.

[102] Кириллов А. А. Пределы. — М.: Наука, 1968.

[103] Маркушевич А. И. Ряды. Элементарный очерк. — М.: Наука, 1979.

[104] Маркушевич А. И. Целые функции. Элементарный очерк. — М.: Наука, 1975.

[105] Маркушевич А. И. Краткий курс теории аналитических функций. — М.: Наука, 1966.

[106] Маркушевич А. И. Замечательные синусы. Введение в теорию эллиптических функций. — М.: Наука, 1965.

[107] Понтрягин Л. С. Знакомство с высшей математикой: Метод координат. — М.: Наука, 1987.

[108] Понтрягин Л. С. Знакомство с высшей математикой: Дифференциальные уравнения и их приложения. — М.: Наука, 1988.

[109] Понтрягин Л. С. Математический анализ для школьников. — М.: Наука, 1988.

[110] Спивак М. Математический анализ на многообразиях. — М.: Мир, 1968.

[111] Тихомиров В. М. Рассказы о максимумах и минимумах. — М.: Наука, 1986.

Нам хотелось бы также отметить несколько книг по такому нетрадиционному для элементарной математики разделу, как **теория особенностей**.

[112] Арнольд В. И. Теория катастроф. — М.: Наука, 1990.

[113] Брус Дж., Джиблин П. Кривые и особенности. — М.: Мир, 1988

[114] Постон Т., Стюарт И. Теория катастроф и ее приложения. — М.: Мир, 1980

[115] Стюарт И. Тайны катастрофы. — М.: Мир, 1987.

Теория вероятностей

[116] Гнеденко Б. В., Хинчин А. Я. Элементарное введение в теорию вероятностей. — М.: Наука, 1982.

[117] Колмогоров А. Н., Журбенко И. Г., Прохоров А. В. Введение в теорию вероятностей. — М.: Наука, 1982.

[118] Мостеллер Ф. Пятьдесят занимательных вероятностных задач с решениями. — М.: Наука, 1985.

Предметный указатель

З

И

К

П

Рихард Курант, Герберт Роббинс

ЧТО ТАКОЕ МАТЕМАТИКА?

Тех. редактор В. Радионов
Корректор А. Ю. Котова
Подготовка иллюстраций В. Крюков

Издательство Московского Центра
непрерывного математического образования

Лицензия ИД №01335 от 24.03.2000 г.

Подписано в печать 30.03.2001 г. Формат 60×90 $^1/_{16}$. Бумага офсетная №1.
Печать офсетная. Печ. л. 35,5. Тираж 3000 экз. Заказ №

МЦНМО
121002, Москва, Большой Власьевский пер., 11

Отпечатано с готовых диапозитивов в ОАО «Типография „Новости“»
107005, Москва, ул. Фридриха Энгельса, 46